G H Patel College of Engineering & Technology

(A Constituent College of Charutar Vidya

Mandal University) V.V.Nagar

DEPARTMENT OF INFORMATION TECHNOLOGY

Mini Project Report

on

Breast Cancer Detection using Machine Learning

Submitted By

Name of Student: Dhruvi Shah, Dhwani Shah, Greena Patel

Enrollment Number: 12202080501011, 12202080501013, 12202080501016

Guided By

Dr. Miral Patel

MINI PROJECT (202040601)

A.Y. 2024-25 EVEN TERM

CERTIFICATE

This is to certify that the Mini Project Report submitted entitled “**Breast Cancer Detection using Machine Learning**” has been carried out by **Dhruvi Shah, Dhwani Shah, Greena Patel** (12202080501011, 12202080501013, 12202080501016) under guidance in partial fulfillment for the Degree of Bachelor of Engineering in Information Technology, 6th Semester of G H Patel College of Engineering & Technology, CVM University, Vallabh Vidyanagar during the academic year 2024-25.

Dr. Miral Patel
Internal Guide

Dr. Nikhil Gondaliya
Head of Department

DECLARATION

I **Dhruvi Shah (12202080501011)**, hereby declare that this Mini Project Report submitted in partial fulfillment for the degree of Bachelor of Engineering in **Information Technology , GCET**. The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar, is a bonafide record of work carried out by me under the supervision of **Dr. Miral Patel** and that no part of this report has been directly copied from any student's reports or taken from any other source, without providing due reference.

Name of the Student   Sign of Student

_____________________________   _____________________________

Date :

Place : Vallabh Vidyanagar

DECLARATION

I **Dhwani Shah(12202080501013)**, hereby declare that this Mini Project Report submitted in partial fulfillment for the degree of Bachelor of Engineering in **Information Technology , GCET**. The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar, is a bonafide record of work carried out by me under the supervision of **Dr. Miral Patel** and that no part of this report has been directly copied from any student's reports or taken from any other source, without providing due reference.

Name of the Student

Sign of Student

Date :

Place : Vallabh Vidyanagar

GCET
G.H. PATEL COLLEGE OF
ENGINEERING & TECHNOLOGY

CVM
UNIVERSITY

DECLARATION

I **Greena Patel (12202080501016)**, hereby declare that this Mini Project Report submitted in partial fulfillment for the degree of Bachelor of Engineering in **Information Technology , GCET**. The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar, is a bonafide record of work carried out by me under the supervision of **Dr. Miral Patel** and that no part of this report has been directly copied from any student's reports or taken from any other source, without providing due reference.

Name of the Student

Sign of Student

Date :

Place : Vallabh Vidyanagar

ACKNOWLEDGEMENT

The completion of any seminar work depends upon cooperation, co-ordination and combined efforts of several sources of knowledge. I would like to express our deepest thanks to **Dr. Miral Patel**, for their valuable inputs, guidance, encouragement, wholehearted cooperation and constructive criticism throughout the duration of my project. I hope that this mini project work report will provide all necessary information required to readers to fulfil their aspiration. Man's quest for knowledge never ends. Theory and practices are essential and complementary to each other.

Dhruvi Shah	Dhwani Shah	Greena Patel
(12202080501011)	(12202080501013)	(12202080501016)

12202080501011
12202080501013
12202080501016

ABSTRACT

Breast cancer is the most commonly diagnosed cancer among women worldwide and poses a major threat to global public health. This project explores the use of machine learning algorithms to automate the detection and classification of breast tumors as either benign or malignant, using the Wisconsin Breast Cancer Diagnostic dataset. The study compares three supervised classification algorithms: **Random Forest**, **K-Nearest Neighbors (KNN)**, and **XGBoost**. Through data preprocessing, feature scaling, and hyperparameter tuning, each model is trained and evaluated based on metrics like **accuracy**, **precision**, **recall**, **F1-score**, and **confusion matrix**. The results demonstrate that the XGBoost model outperforms others, offering a robust and scalable solution for aiding early cancer detection.

List of Figures

Figure 1: Overall Flow of Project.. 2

Figure 2: Model Evaluation of 70-30% Split... 16

Figure 3: Visualization of Result 70-30% Split 17

Figure 4: Model Evaluation of 80-20% Split...18

Figure 5: Visualization of Result 80-20% Split 19

Figure 6: Model Evaluation of 90-10% Split... 20

Figure 7: Visualization of Result 90-10% Split 21

Figure 8: Flowchart 23

Figure 9:Gantt Chart.. 24

List of Tables

Table 1: Limitation of existing system.. 5

Table 2: Advantages of proposed system...5

Table 3: Hardware Requirements for the project.. 6

Table 4: Software Requirements for the project... 6

Table 5: Technical Feasibility...7

Table 6: Economical Feasibility 7

Table 7: Operational Feasibility 7

Table 8: Output..8

Table 9:Model evaluation of 70-30% Split 16

Table 10: Model evaluation of 80-20% Split 18

Table 11: Model evaluation of 90-10% Split..20

12202080501011
12202080501013
12202080501016

Table of Contents

Chapter-1: Introduction...1

1.1 Problem Statement...1

1.2 Project Overview 1

1.3 Objective of the project 2

1.4 Research Objectives...3

1.5 Scope and Significance 3

1.6 Open Research Issues 4

1.6.1 Dataset Limitations 4
1.6.2 Model Generalization 4
1.6.3 Real World Integration...4
1.6.4 Interpretability and Trust 4
1.6.5 Ethical and Legal Concerns ………………………………………….4

Chapter-2: System analysis 5

2.1 Existing System Analysis 5

2.2 Proposed System Analysis...5

2.3 System Requirements 6

2.4 Feasibility Analysis...6

2.4.1 Technical Feasibility...6
2.4.2 Economic Feasibility...6
2.4.3 Operational Feasibility 6

2.5 Input and Output Analysis...7

Chapter-3: Literature Review 9

3.1Evolution of Breast Cancer Detection using Machine Learning 9

3.2 Machine Learning algorithms for Breast Cancer Detection 9

3.3 Key Studies and Findings……………………………………………10

3.4 Challenges and Criticism ……………………………………………10

3.5 Research Trends and Future Directions………………………………11

3.6 Summary………………………………………………………………..12

Chapter-4: Methodology & Implementation ....…13

4.1 Dataset Description and Preprocessing...13

4.1.1 Dataset Overview 13

4.1.2 Data Cleaning and Preprocessing 13
4.2 Implementation...13
4.3 Machine Learning Models and Hyperparameter Tuning 14
4.3.1 Random Forest Classifier…………………………………………14
4.3.2 K-Nearest Neighbors……………………………………………...14
4.3.3 Extreme Gradient Boosting(XGBoost)…………………………15
4.4 Model Training and Evaluation………………………………………16
4.4.1 Evaluation Workflow……………………………………………16
4.4.2 Model Evaluations………………………………………….... 16
Chapter-5: System Design 23
5.1 Flowchart 23
5.2 Gantt Chart 25
Chapter 6: Conclusion and Future Work 26
6.1 Conclusion..26
6.2 Future Work 26
Chapter-7: References...…28

Chapter 1: Introduction

1.1 Problem Statement

Breast cancer is one of the leading causes of cancer-related deaths among women globally. Early and accurate diagnosis is critical for improving patient outcomes and survival rates. Traditionally, diagnosis is conducted through physical examinations, mammography, and biopsies followed by histopathological analysis. While effective, these methods are time-intensive, require substantial medical expertise, and are prone to diagnostic subjectivity and inter-observer variability.

In recent years, the rise of digitized healthcare data has created new opportunities for leveraging artificial intelligence and machine learning (ML) to aid in medical diagnosis. ML models can analyze large datasets to identify complex patterns and correlations that may not be immediately evident through manual analysis. These models, when properly trained and validated, offer a faster, more consistent, and scalable approach to diagnosing breast cancer, especially in areas with limited access to experienced medical professionals or diagnostic infrastructure.

This project seeks to address the growing need for intelligent, automated diagnostic systems by developing a machine learning-based breast cancer detection model that classifies tumors as benign or malignant based on diagnostic feature inputs.

1.2 Project Overview

The proposed project implements a comparative analysis of three machine learning classification algorithms—Random Forest, K-Nearest Neighbors (KNN), and XGBoost—on the Wisconsin Breast Cancer Diagnostic dataset. The dataset contains 30 numeric features extracted from digitized cell images of breast mass tissue obtained via fine needle aspirate (FNA).

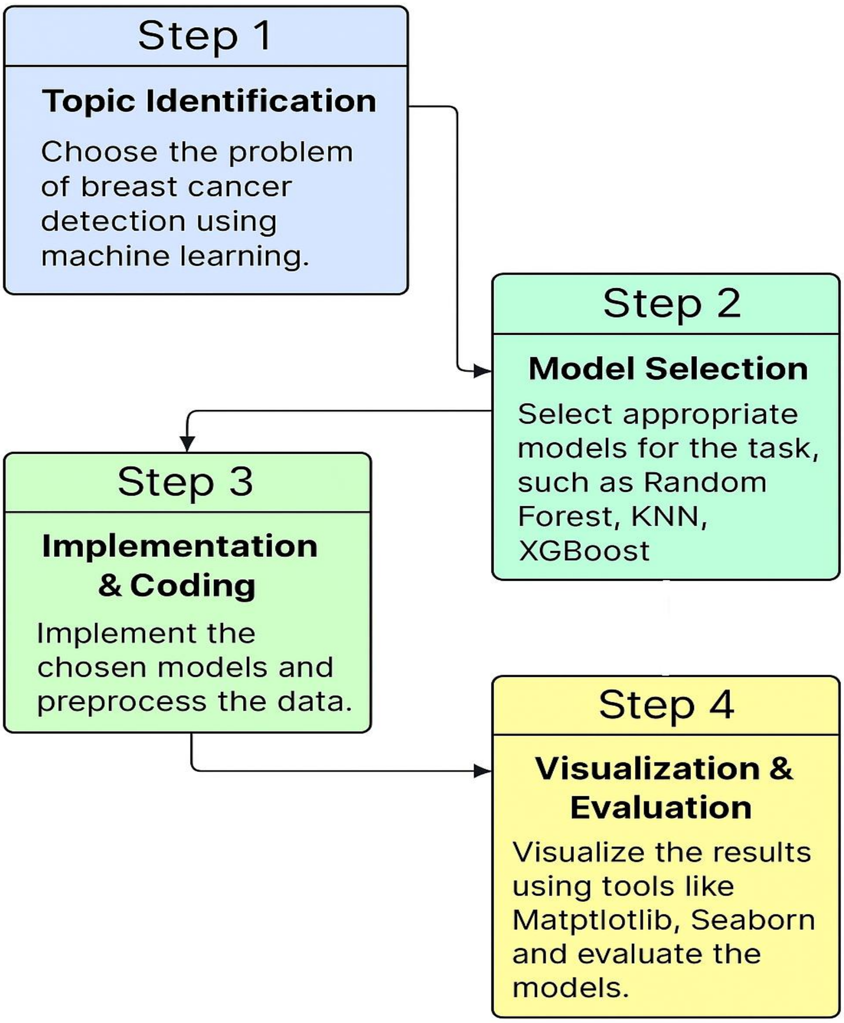
The workflow involves:

- Cleaning and preprocessing the dataset,
- Feature scaling to normalize input attributes,
- Splitting the data into training and testing sets,
- Training each classifier,
- Evaluating their performance using statistical metrics, and
- Recommending the most accurate and reliable model for potential real-world deployment in healthcare applications.

1.3 Objectives

The specific objectives of this project are:

- To preprocess the Wisconsin Breast Cancer dataset by eliminating irrelevant columns, managing missing values, and standardizing features.
- To implement and compare the classification performance of Random Forest, KNN, and XGBoost models.
- To evaluate each model using accuracy, precision, recall, F1-score, and confusion matrix.
- To visualize classification results and model comparisons using plots and heatmaps.
- To identify the most accurate model for potential integration into diagnostic support systems.


Step 1
Topic Identification
Choose the problem of breast cancer detection using machine learning.
Step 2
Model Selection
Select appropriate models for the task, such as Random Forest, KNN, XGBoost
Step 3
Implementation & Coding
Implement the chosen models and preprocess the data.
Step 4
Visualization & Evaluation
Visualize the results using tools like Matptlotlib, Seaborn and evaluate the models.

Overall Flow of the Project

1.4 Research Objectives

The research component of this project is directed at answering the following core questions:

- How accurately can machine learning algorithms classify breast cancer as benign or malignant using numerical diagnostic features?
- Which algorithm among Random Forest, KNN, and XGBoost provides the best balance of precision, recall, and generalization?
- How do data preprocessing techniques such as feature scaling and encoding influence the model performance?
- What are the practical limitations and considerations for deploying such a system in clinical environments?

1.5 Scope and Significance

Scope:

- The scope of this study is limited to the use of structured tabular data (not image-based) for breast cancer detection.
- Only three machine learning classifiers are considered in this project: Random Forest, KNN, and XGBoost.
- The model is trained and tested on a publicly available dataset (Wisconsin Breast Cancer Diagnostic), with no real-time data integration or hardware deployment.

Significance:

- Early Detection: The proposed system can assist in the early detection of breast cancer, increasing the likelihood of successful treatment.
- Decision Support: It can serve as a decision support tool for radiologists and oncologists by providing a second opinion.
- Automation: Automation can help reduce the workload on medical professionals and minimize diagnostic errors.
- Accessibility: This solution can be adapted to serve rural or under-resourced regions where expert diagnostic services are limited.
- Foundation for Future Research: The results can be used to build more complex systems integrating image data, real-time diagnostics, and patient history.

1.6 Open Research Issues

Despite the promising results and potential of machine learning in cancer detection, several challenges remain:

1.6.1 Dataset Limitations:

- Public datasets like the Wisconsin Diagnostic dataset are well-structured but may not reflect real-world diversity such as noise, missing data, or unbalanced class distribution.
- Most available datasets do not include demographic or clinical variables, which could improve prediction performance.

1.6.2 Model Generalization:

- Machine learning models trained on one dataset may not generalize well to other datasets or clinical populations.
- Overfitting on small datasets is a common issue, particularly with high-complexity models like XGBoost.

1.6.3 Real-World Integration:

- Integrating ML-based diagnostic tools into hospital systems requires compliance with data privacy regulations (like HIPAA).
- Real-world deployment also requires continuous model updating and validation using recent patient data.

1.6.4 Interpretability and Trust:

- Healthcare professionals require explainable AI systems. Black-box models like Random Forest and XGBoost, although accurate, often lack transparency.
- Building trust in AI-driven diagnosis tools is essential for adoption in clinical practice.

1.6.5 Ethical and Legal Concerns:

- Accountability in the case of incorrect predictions must be clearly defined.
- Patient consent and ethical handling of medical data remain major considerations.

Chapter 2: System analysis

2.1 Existing System Analysis

In the conventional breast cancer diagnosis process, biopsy samples are manually examined under a microscope by pathologists. While this method is accurate in expert hands, it is highly dependent on human skill and is subject to interpretation errors, time delays, and limited scalability in rural or under-resourced settings.

Limitations of the Existing System

Feature	**Manual Diagnosis System**
Accuracy	Relies heavily on expert interpretation
Speed	Time-consuming diagnostic process
Cost	High (equipment and human expertise)
Consistency	Varies with observer experience
Accessibility	Limited in rural/remote healthcare areas
Scalability	Not scalable due to need for specialists

Table 1

2.2 Proposed System Analysis

The proposed system uses a **machine learning-based approach** to automate breast cancer detection using the **Breast Cancer Wisconsin (Diagnostic) dataset**. It implements three classifiers—**Random Forest**, **K-Nearest Neighbors (KNN)**, and **XGBoost**—to classify tumors as benign or malignant based on numerical features extracted from FNA images.

The system was built and executed on **Windows OS using Google Colab**, with **Python** and libraries like **scikit-learn**, **XGBoost**, **Matplotlib**, and **Seaborn**.

Advantages of the Proposed System

Feature	**ML-Based Diagnosis System**
Accuracy	High (Up to 97.66% with XGBoost)
Speed	Real-time or near-instant predictions
Cost	Low operational cost post-deployment
Consistency	Consistent results from trained models

Feature	ML-Based Diagnosis System
Accessibility	Can be deployed on cloud/web/mobile
Scalability	Easily scalable with minimal infrastructure

Table 2

2.3 System Requirements

Hardware Requirements

Component	Specification
Operating System	Windows 10 or higher
RAM	Minimum 4 GB (Recommended: 8 GB)
Storage	At least 2 GB free disk space
Internet	Required for Google Colab access

Table 3

Software Requirements

Software Tool	Purpose
Python 3.x	Programming language
scikit-learn	Machine learning models
XGBoost	Gradient boosting model
Google Colab	Cloud-based coding environment
Matplotlib, Seaborn	Visualization libraries

Table 4

2.4 Feasibility Analysis

1. Technical Feasibility

- Readily available ML frameworks like **scikit-learn** and **XGBoost**.
- Executed on cloud (Google Colab) without high-end hardware needs.
- Compatible with various datasets for future expansion.

Factor	Description
Tools	Python, Colab, ML libraries
Training Time	Moderate (~few minutes per model)
Deployment	Feasible on web or mobile platforms

Table 5

2. Economic Feasibility

- Free tools and platforms (Python, Colab).
- No cost for data as the dataset is open-source.
- Long-term cost savings from automation and reduced human involvement.

Expense Area	Cost Evaluation
Tools	Free (open-source)
Infrastructure	Low (cloud-based, no local GPU required)
Maintenance	Minimal after initial development

Table 6

3. Operational Feasibility

- The system is user-friendly and can run on any modern browser with internet access.
- Easily interpretable outputs (e.g., prediction labels and graphs).
- Can be operated by medical assistants or patients with basic technical understanding.

Factor	Description
Usability	High – simple interface via notebooks/web
Integration	Can be embedded in healthcare platforms
User Training	Minimal or none required

Table 7

2.5 Input and Output Analysis

Input

- Cleaned and scaled dataset with 30 diagnostic numerical features extracted from FNA of breast tissue.
- Input format: CSV file.

Output

- Classification result: **Benign (0)** or **Malignant (1)**
- Evaluation metrics: Accuracy, Precision, Recall, F1-Score
- Visual outputs: Confusion matrix, metric comparison bar charts

Category	**Description**
Input	Preprocessed CSV (30 features)
Output	Model prediction (0 or 1), reports, graphs
Tools for Output	Matplotlib, Seaborn

Table 8

Chapter 3 : Literature Review

3.1 Evolution of Breast Cancer Detection Using Machine Learning

- **Early Phase (Pre-2000s):** Initial approaches relied on statistical models and manual feature selection, which limited generalizability. Models such as logistic regression and decision trees were used but often lacked the ability to handle complex patterns in data.
- **Introduction of Machine Learning (2000–2010):** The advent of basic ML algorithms such as **KNN, Decision Trees, and SVM** led to improved accuracy in pattern recognition. These models were applied to structured datasets like the **Wisconsin Breast Cancer Dataset (WBCD)**, demonstrating better classification performance.
- **Rise of Deep Learning (2010–2015):** With the introduction of **Convolutional Neural Networks (CNNs)**, feature extraction and classification tasks became automated, significantly improving accuracy in breast cancer classification, particularly with histopathological and mammographic images.
- **Advanced Ensemble Models (2015–2020):** Ensemble methods, including **Random Forest, XGBoost, and AdaBoost**, emerged as powerful tools by combining multiple classifiers to enhance predictive performance. Studies highlighted the superior performance of these models in handling imbalanced data and improving classification accuracy.
- **Modern Era (2020–Present):** Current approaches focus on **multi-modal and personalized models** that integrate imaging, clinical, and genomic data. Additionally, **transfer learning** and **federated learning** techniques enhance model generalizability, privacy, and security across different healthcare institutions.

3.2 Machine Learning Algorithms for Breast Cancer Detection

- **K-Nearest Neighbors (KNN)**

  KNN classifies data points based on similarity to their nearest neighbors. It is a simple yet effective model, particularly for small datasets. However, it is sensitive to noise and suffers from degraded performance with high-dimensional data. Despite these limitations, KNN has been widely used in breast cancer detection due to its ease of implementation and reasonable classification performance.

- **Random Forest (RF)**

  RF is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions to improve classification accuracy. It is known for its robustness in handling imbalanced datasets and high-dimensional data. RF has been extensively used in breast cancer detection, with studies reporting accuracy rates often exceeding 95%. Its ability to handle large datasets and identify feature importance makes it a preferred choice for classification tasks.

- **Extreme Gradient Boosting (XGBoost)**

  XGBoost is a gradient-boosting algorithm that optimizes classification by sequentially reducing errors. It uses advanced regularization techniques to prevent overfitting and improve computational efficiency. Studies have demonstrated that XGBoost achieves higher precision, recall, and F1-scores compared to traditional models, making it ideal for breast cancer prediction.

3.3 Key Studies and Findings

- **Mahmood et al. (2025):** RF achieved 94% accuracy and 100% sensitivity in a South Iraq study that analyzed 415 instances from Al-Sadr Teaching Hospital.
- **Al-Nawashi et al. (2024):** CNN-based feature extraction combined with RF and SVM models improved classification accuracy on 3002 mammogram images.
- **Akash et al. (2024):** RF and Logistic Regression achieved 96.49% and 98.25% accuracy, respectively, in classifying breast biopsy samples, emphasizing the importance of feature selection and hyperparameter tuning.
- **Bhise et al. (2023):** CNN and RF models outperformed KNN in classifying BreaKHis dataset images, demonstrating the effectiveness of hybrid models in breast cancer classification.
- **Sharma et al. (2018)**: Compared Random Forest, k-Nearest Neighbors (kNN), and Naïve Bayes classifiers using the Wisconsin Diagnosis Breast Cancer dataset. kNN achieved the highest accuracy of 95.90%, outperforming other models.
- **Al-Imran et al. (2024)**: Conducted a comparative analysis of ML algorithms, finding that Random Forest and SVM achieved superior predictive performance on the Wisconsin Breast Cancer dataset.
- **Dhahri et al. (2019)**: Utilized genetic programming for feature selection and parameter optimization, significantly improving breast cancer classification accuracy.

3.4 Challenges and Criticisms

- **Lack of Model Interpretability**
  Deep learning models often operate as “black boxes,” making it difficult for clinicians to understand how predictions are made. This lack of transparency limits the trust and clinical adoption of AI-based diagnostic systems.

- **Class Imbalance and Data Bias**
  Breast cancer datasets often contain more benign cases than malignant ones, leading to biased predictions and higher false-negative rates. Techniques such as **Synthetic Minority Over-sampling Technique (SMOTE)** and cost-sensitive learning are commonly applied to mitigate this issue.

- **Overfitting and Poor Generalization**

ML models trained on small or homogeneous datasets may overfit, leading to poor performance on unseen data. Cross-validation, dropout regularization, and transfer learning techniques are used to enhance model generalizability.

- **High Computational Costs**
  Deep learning models, especially CNNs, require significant computational resources, making them less accessible to resource-constrained healthcare institutions. Optimizing model architecture and leveraging cloud-based solutions can help address this issue.

- **Privacy and Ethical Concerns**
  Handling sensitive medical data raises concerns about patient privacy and data security. **Federated learning** is emerging as a promising solution that enables collaborative model training without compromising patient privacy.

3.5 Recent Trends and Future Directions

- **Explainable AI (XAI)**
  Explainable AI (XAI) models such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are being adopted to enhance the interpretability and trustworthiness of AI models in healthcare.

- **Multi-Modal and Personalized Models**
  Integrating diverse data modalities, including imaging, clinical, and genomic data, enhances model accuracy and enables personalized treatment plans. Personalized AI models have the potential to improve patient outcomes by tailoring treatment recommendations based on individual risk factors.

- **Federated Learning for Privacy**
  Federated learning allows multiple institutions to collaboratively train models without sharing sensitive patient data. This approach ensures data privacy while enhancing model robustness and generalizability.

- **Real-Time and Cloud-Based Diagnostic Systems**
  Cloud-based platforms facilitate real-time breast cancer diagnosis and seamless integration into clinical workflows. These systems provide faster results and improve access to quality healthcare, especially in remote areas.

3.6 Summary

Machine learning has revolutionized breast cancer detection by automating feature extraction and improving classification accuracy. While models such as KNN, RF, and XGBoost have shown impressive performance, challenges such as model interpretability, class imbalance, and privacy concerns remain. Moving forward, integrating explainable AI, multi-modal data, and federated learning holds the promise of enhancing the accuracy, interpretability, and clinical adoption of breast cancer detection models. Addressing these challenges will enable the development of more reliable and efficient AI-powered diagnostic systems.

Chapter 4: Methodology and Implementation

4.1 Dataset Description and Preprocessing

4.1.1 Dataset Overview

The dataset used in this study is the **Breast Cancer Wisconsin Diagnostic dataset**, which contains detailed numerical attributes of tumors. Each sample is classified as either **malignant (M)** or **benign (B)** based on fine needle aspiration biopsy results. The dataset comprises 30 numerical features extracted from cell nuclei, representing characteristics such as texture, perimeter, smoothness, symmetry, and fractal dimensions.

4.1.2 Data Cleaning and Preprocessing

To ensure high-quality data for training and evaluation, the following preprocessing steps were applied:

1. **Feature Selection and Removal of Irrelevant Columns**: Non-informative columns such as 'id' and 'Unnamed: 32' (if present) were removed to avoid redundancy and improve model efficiency.
2. **Handling Missing Values**: Any missing values were identified and removed to maintain dataset integrity.
3. **Encoding Categorical Variables**: The target variable 'diagnosis' was converted into numerical format, where **M = 1 (Malignant)** and **B = 0 (Benign)** to facilitate binary classification.
4. **Feature Scaling (Standardization)**: The **StandardScaler** was applied to normalize all numerical features, ensuring a mean of zero and unit variance. This step enhances model performance, particularly for algorithms sensitive to feature magnitudes, such as KNN and gradient boosting methods.

4.2 Data Splitting Strategy

To evaluate model generalizability, multiple **train-test splits** were utilized:

- **70:30 Split**: 70% training, 30% testing.
- **80:20 Split**: 80% training, 20% testing.
- **90:10 Split**: 90% training, 10% testing.

Additionally, for **XGBoost**, an extra **10% of the training data** was set aside as a validation set to facilitate early stopping and prevent overfitting.

4.3 Machine Learning Models and Hyperparameter Tuning

This study assessed the performance of three machine learning models: **Random Forest (RF), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost).**

4.3.1 Random Forest Classifier (RF)

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to enhance robustness and reduce overfitting. It operates by training a collection of decision trees on different subsets of the data and averaging their predictions to improve accuracy and stability. The randomness introduced in feature selection and data sampling helps mitigate overfitting, making RF a powerful choice for classification tasks.

This algorithm is particularly effective in handling high-dimensional datasets and capturing complex feature interactions without requiring extensive data preprocessing. Furthermore, RF is resilient to outliers and missing data, making it suitable for medical diagnostics applications.

Hyperparameter Tuning

GridSearchCV with **5-fold cross-validation** was employed to optimize the following parameters:

- n_estimators: [100, 200, 300] (Number of trees in the forest)
- max_depth: [5, 10, None] (Maximum depth of trees)
- min_samples_split: [2, 5, 10] (Minimum samples required to split a node)

4.3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a crucial algorithm utilized in this study, recognized for its straightforwardness and effectiveness in handling classification tasks. KNN is based on the concept of instance-based learning, where the classification of a new data point is decided by the majority vote of its closest labeled neighbors. The proximity of these neighbors is typically assessed using distance metrics such as Euclidean distance.

This algorithm's simplicity makes it highly intuitive, as it relies directly on the nearest examples in the dataset to make predictions. This characteristic is particularly advantageous when dealing with non-linear decision boundaries, as KNN can adapt to the underlying structure of the data without needing complex assumptions.

One of the key benefits of KNN is its ease of implementation and interpretation, which makes it a valuable tool in exploratory data analysis and in the initial stages of model development. Despite its straightforward nature, KNN can still achieve competitive performance, especially when the data is evenly distributed and the features are relevant and informative. This combination of simplicity and effectiveness makes KNN a versatile choice in classification tasks.

Hyperparameter Tuning

- n_neighbors: [3, 5, 7, 9] (Number of neighbors considered for classification)
- weights: ['uniform', 'distance'] (Uniform assigns equal weights; Distance assigns higher weights to closer neighbors)

GridSearchCV with **5-fold cross-validation** was utilized to determine the best combination of parameters.

4.3.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a powerful and efficient gradient boosting algorithm that iteratively improves classification performance by minimizing errors. It builds sequential decision trees, where each new tree corrects the residual errors of the previous trees. XGBoost incorporates regularization techniques to prevent overfitting, making it particularly well-suited for complex classification problems.

XGBoost is designed to handle missing values automatically and is optimized for both speed and memory efficiency. Its ability to efficiently process large datasets and capture intricate feature interactions makes it one of the top choices for structured data classification tasks.

Hyperparameter Tuning

- n_estimators: 1000 (Number of boosting rounds)
- learning_rate: 0.01 (Step size for weight updates)
- max_depth: 6 (Maximum depth of trees)
- subsample: 0.9 (Fraction of samples used per boosting round)
- colsample_bytree: 0.9 (Fraction of features used per tree)
- early_stopping_rounds: 20 (Stops training if no improvement is observed after 20 iterations)

Early stopping was incorporated during training to enhance model generalization.

4.4 Model Training and Evaluation

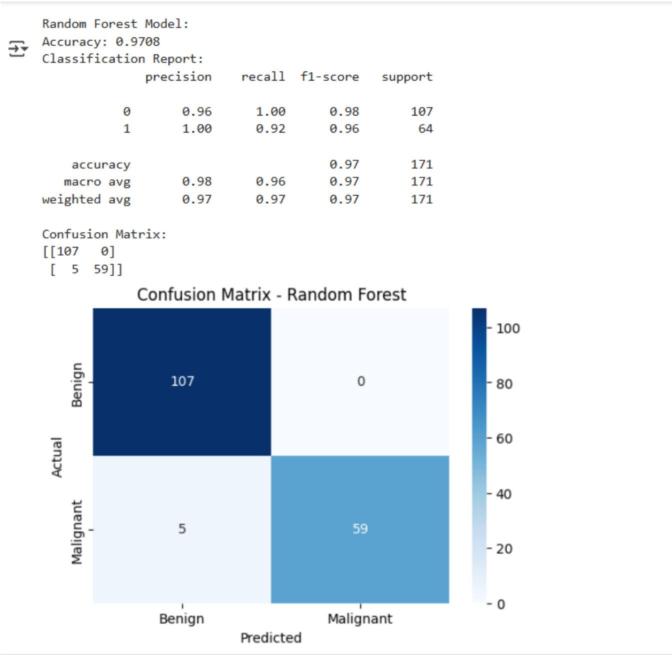
Each model was trained using the respective training set and evaluated on the test set using multiple performance metrics:

- **Accuracy**: Measures the proportion of correctly predicted instances.
- **Precision**: The fraction of correctly identified positive cases among predicted positives.
- **Recall (Sensitivity)**: The fraction of actual positive cases correctly identified.
- **F1-score**: The harmonic mean of precision and recall, balancing false positives and false negatives.
- **Confusion Matrix**: A diagnostic tool to visualize classification performance in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

4.4.1 Evaluation Workflow

1. Train each model on the designated training set.
2. Test the trained model on unseen test data.
3. Generate confusion matrices to analyze classification performance.
4. Compare results across different train-test splits.
5. Select the best-performing model based on accuracy and other key metrics.

4.4.2 Model Evaluation(70-30% training-testing split)

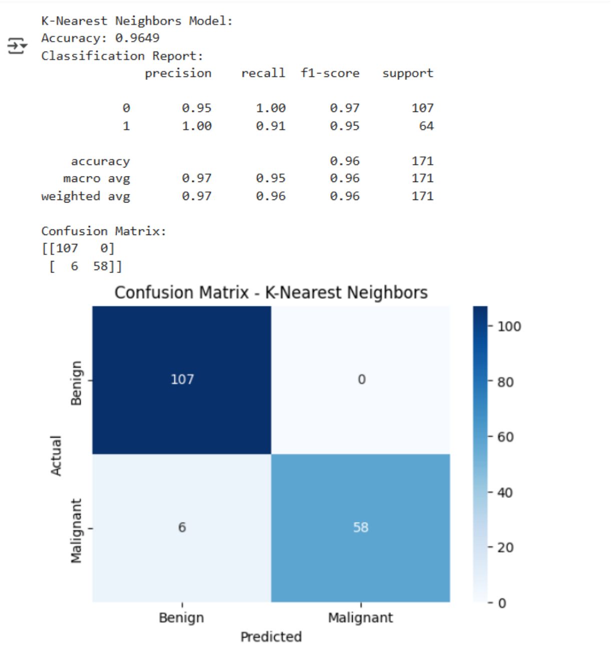


```
Random Forest Model:
Accuracy: 0.9708
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       107
           1       1.00      0.92      0.96        64

    accuracy                           0.97       171
   macro avg       0.98      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171

Confusion Matrix:
[[107   0]
 [  5  59]]
```

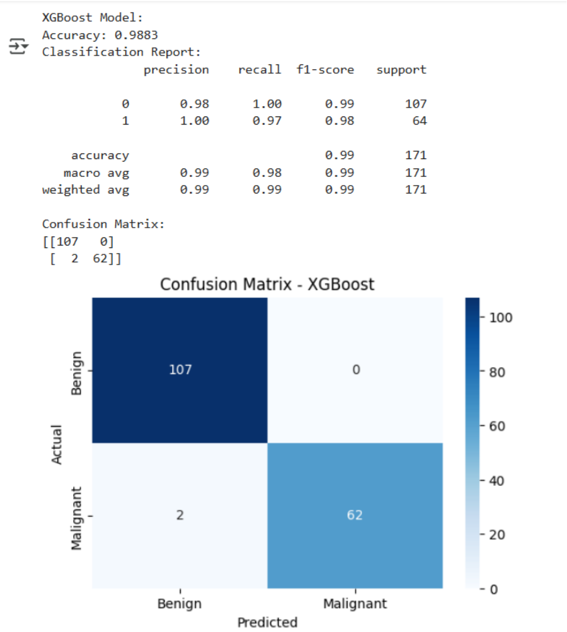


```
K-Nearest Neighbors Model:
Accuracy: 0.9649
Classification Report:
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       107
           1       1.00      0.91      0.95        64

    accuracy                           0.96       171
   macro avg       0.97      0.95      0.96       171
weighted avg       0.97      0.96      0.96       171

Confusion Matrix:
[[107   0]
 [  6  58]]
```



```
XGBoost Model:
Accuracy: 0.9883
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       107
           1       1.00      0.97      0.98        64

    accuracy                           0.99       171
   macro avg       0.99      0.98      0.99       171
weighted avg       0.99      0.99      0.99       171

Confusion Matrix:
[[107   0]
 [  2  62]]
```

Figure 2 : Model Evaluation of 70-30% Split

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	97.08%	97.00%	97.00%	97.00%
K-Nearest Neighbors	96.49%	97.00%	96.00%	96.00%

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	98.83%	99.00%	99.00%	99.00%

Table 9

- Similarly 80-20% and 90-10% training-testing splits are performed**.**

➢ **Visualization of the results**

- Performance Evaluation of Random Forest, KNN, and XG Boost with 70-30 Split

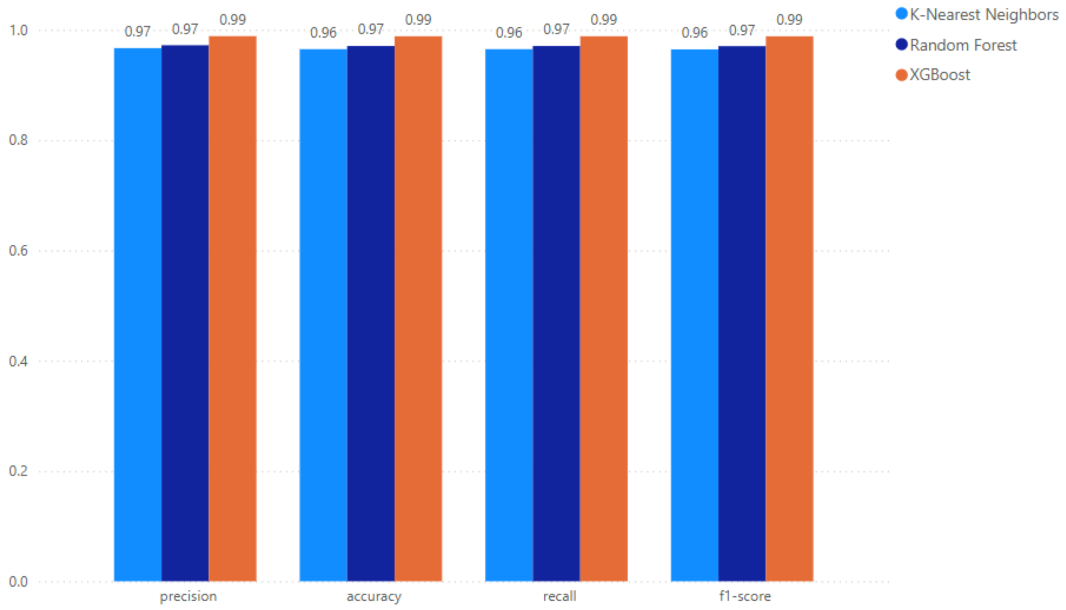
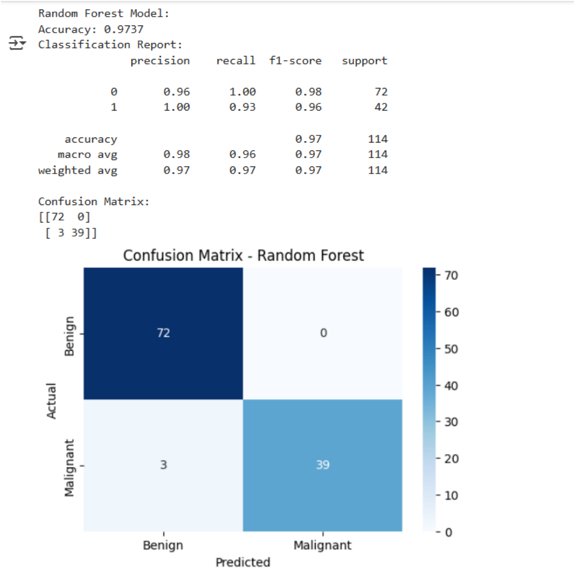


Figure 3 : Visualization of Result 70-30% Split

➢ **Model Evaluation(80-20% training-testing split)**

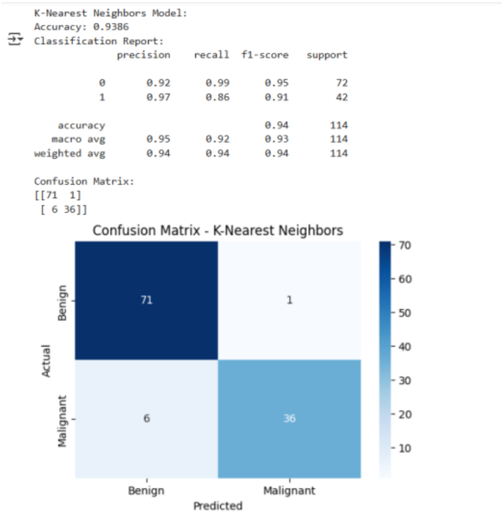


```
Random Forest Model:
Accuracy: 0.9737
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        72
           1       1.00      0.93      0.96        42

    accuracy                           0.97       114
   macro avg       0.98      0.96      0.97       114
weighted avg       0.97      0.97      0.97       114

Confusion Matrix:
[[72  0]
 [ 3 39]]
```

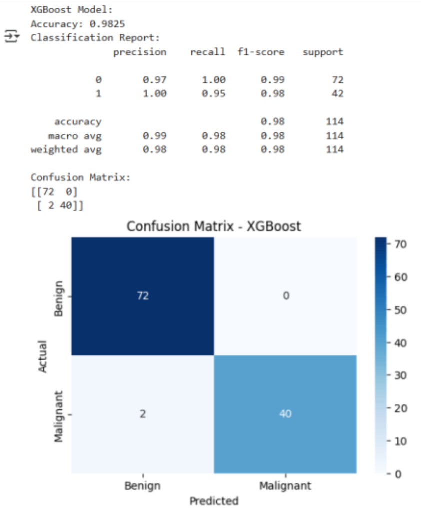


```
K-Nearest Neighbors Model:
Accuracy: 0.9386
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95        72
           1       0.97      0.86      0.91        42

    accuracy                           0.94       114
   macro avg       0.95      0.92      0.93       114
weighted avg       0.94      0.94      0.94       114

Confusion Matrix:
[[71  1]
 [ 6 36]]
```



```
XGBoost Model:
Accuracy: 0.9825
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        72
           1       1.00      0.95      0.98        42

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114

Confusion Matrix:
[[72  0]
 [ 2 40]]
```

Figure 4 : Model Evaluation of 80-20% Split

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	97.08%	98.30%	98.20%	98.24%
K-Nearest Neighbors	96.00%	96.05%	96.00%	96.01%
XGBoost	98.68%	98.70%	98.65%	98.67%

Table 10

➢ Visualization of the results

- Performance Evaluation of Random Forest, KNN, and XG Boost with 80-20 Split

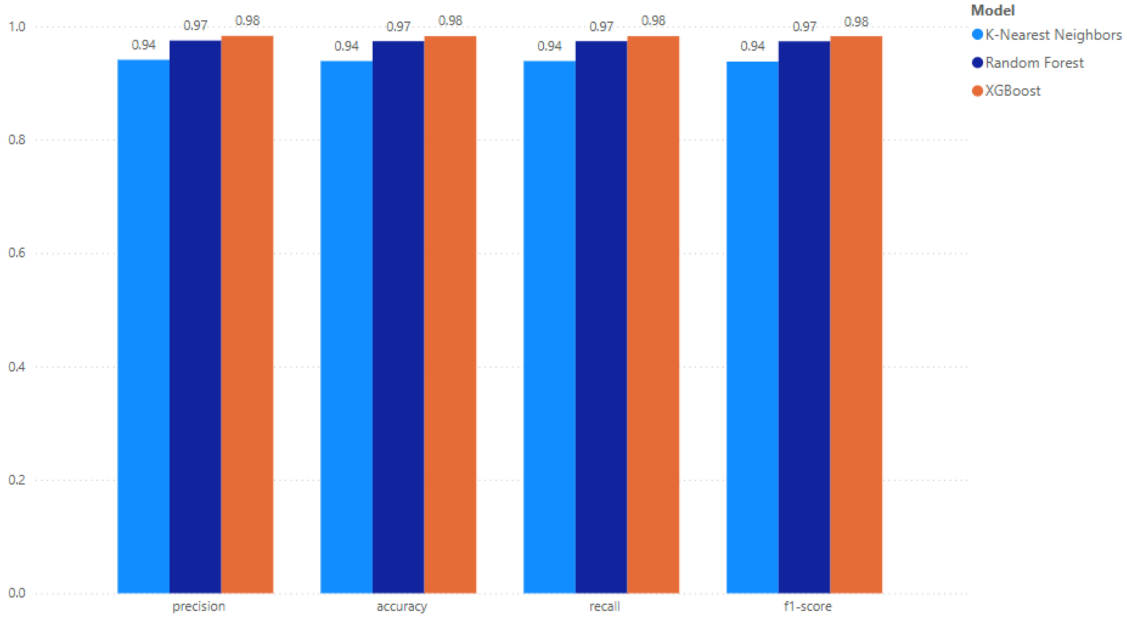
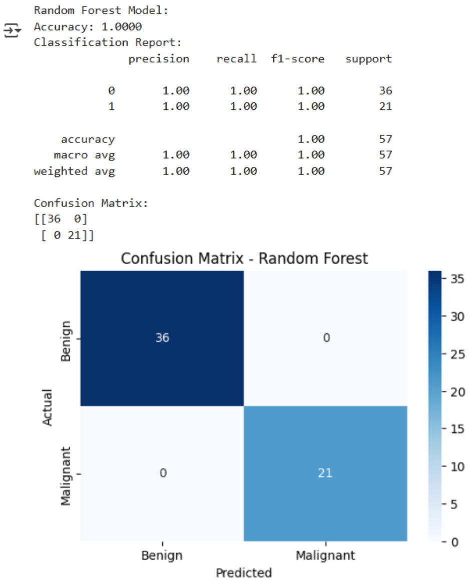


Figure 5 : Visualization of Result 80-20% Split

➢ Model Evaluation(90-10% training-testing split)

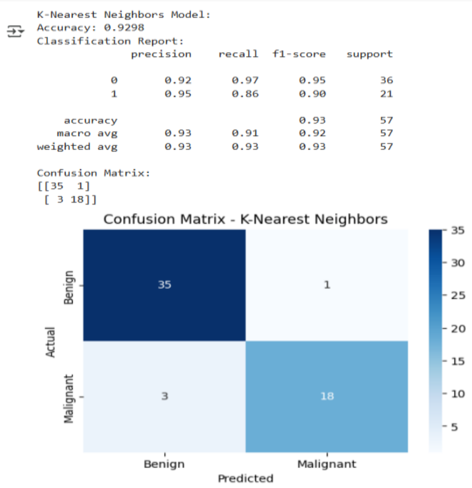


```
Random Forest Model:
Accuracy: 1.0000
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        36
           1       1.00      1.00      1.00        21

    accuracy                           1.00        57
   macro avg       1.00      1.00      1.00        57
weighted avg       1.00      1.00      1.00        57

Confusion Matrix:
[[36  0]
 [ 0 21]]
```

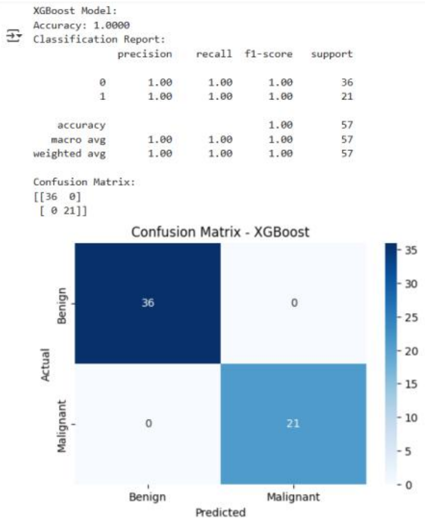


```
K-Nearest Neighbors Model:
Accuracy: 0.9298
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.97      0.95        36
           1       0.95      0.86      0.90        21

    accuracy                           0.93        57
   macro avg       0.93      0.91      0.92        57
weighted avg       0.93      0.93      0.93        57

Confusion Matrix:
[[35  1]
 [ 3 18]]
```



```
XGBoost Model:
Accuracy: 1.0000
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        36
           1       1.00      1.00      1.00        21

    accuracy                           1.00        57
   macro avg       1.00      1.00      1.00        57
weighted avg       1.00      1.00      1.00        57

Confusion Matrix:
[[36  0]
 [ 0 21]]
```

Figure 6 : Model Evaluation of 90-10% Split

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	100.00%	100.00%	100.00%	100.00%
K-Nearest Neighbors	92.98%	93.00%	93.00%	93.00%
XGBoost	100.00%	100.00%	100.00%	100.00%

Table 11

- **Visualization of the results**
- Performance Evaluation of Random Forest, KNN, and XG Boost with 90-10 Split

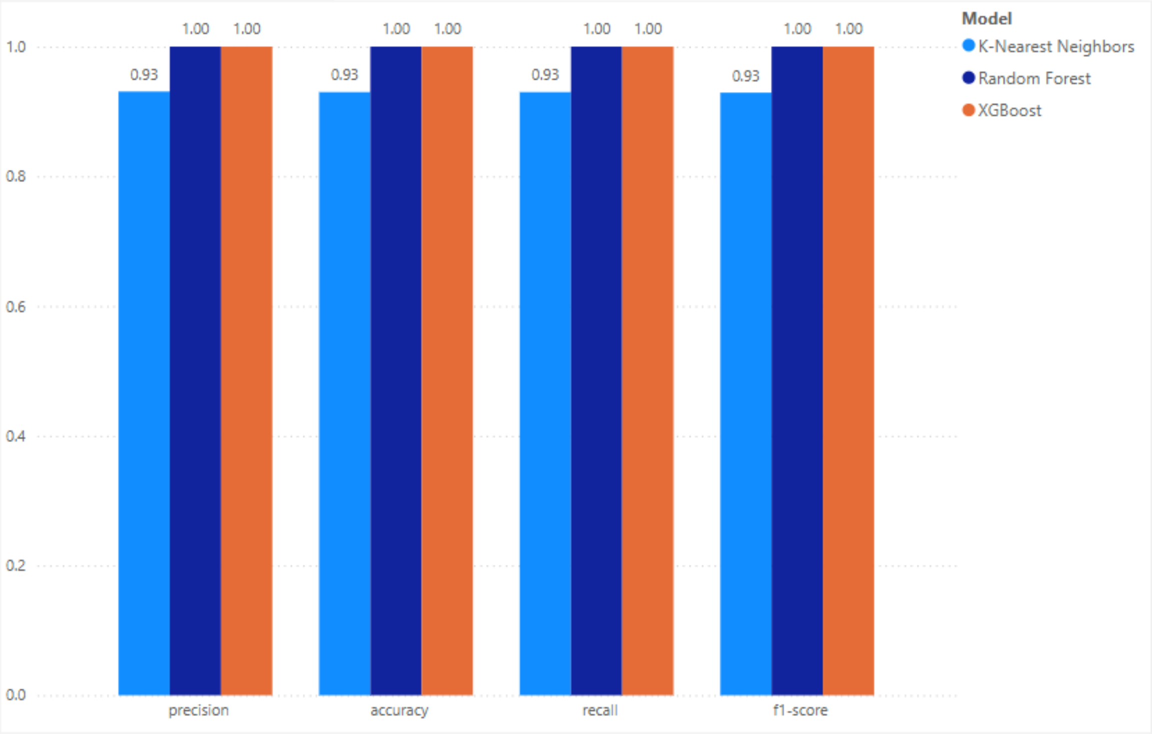


Figure 7 : Visualization of Result 90-10% Split

Chapter 5: System Design

5.1 Flowchart

1. **Start**
   The process begins by defining the objective — detecting breast cancer using ML models.

2. **Data Collection**
   The dataset (Breast_Cancer_Dataset.csv) is loaded into the system. This dataset includes diagnostic measurements of breast cancer tumors and a label (diagnosis) indicating whether the tumor is **malignant (M)** or **benign (B)**.

3. **Data Preprocessing**
   Several steps are performed to prepare the data for modeling:
- Dropping unnecessary columns (like id and unnamed columns).
- Handling missing values.
- Encoding the target variable (diagnosis) as binary (M → 1, B → 0).
- Feature scaling using **StandardScaler**.
- Splitting the data into training and testing sets (90/10 split).

4. **Model Selection**
   Three classification models are selected for evaluation:
- **Random Forest**: An ensemble method using multiple decision trees.
- **K-Nearest Neighbors (KNN)**: A distance-based classification algorithm.
- **XGBoost**: A powerful gradient boosting model.
  Each model is independently tuned and evaluated.

5. **Hyperparameter Tuning**
   For better performance, hyperparameters for each model are optimized using **GridSearchCV**:
- Random Forest: Number of trees, depth, and split rules.
- KNN: Number of neighbors and weighting method.
- XGBoost: Controlled using default parameters, learning rate, depth, etc., with early stopping applied.

6. **Early Stopping (for XGBoost only)**
   To prevent overfitting, **XGBoost** uses early stopping. A portion of the training data is reserved for validation, and training stops when the validation score stops improving.

7. **Model Training**
   Each model is trained on the prepared dataset using the best-found hyperparameters.

8. **Model Evaluation**

The trained models are evaluated on the test set using metrics such as:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **Confusion Matrix** (with heatmaps)

This helps compare performance and select the most accurate model.

9. Stop

The pipeline ends after identifying the best model (e.g., XGBoost, if it has the highest accuracy). The results can then be visualized, saved, or deployed.

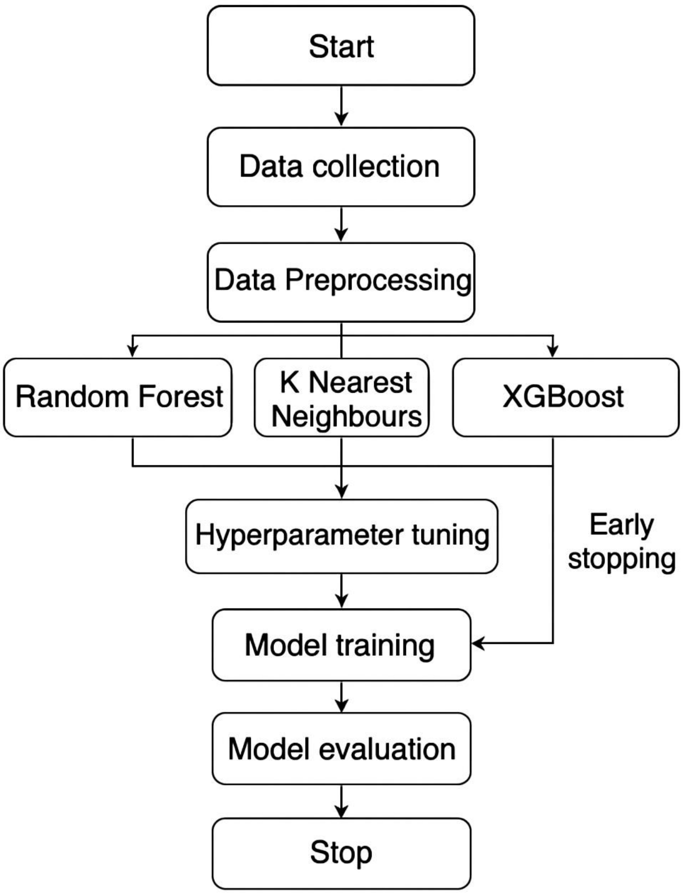
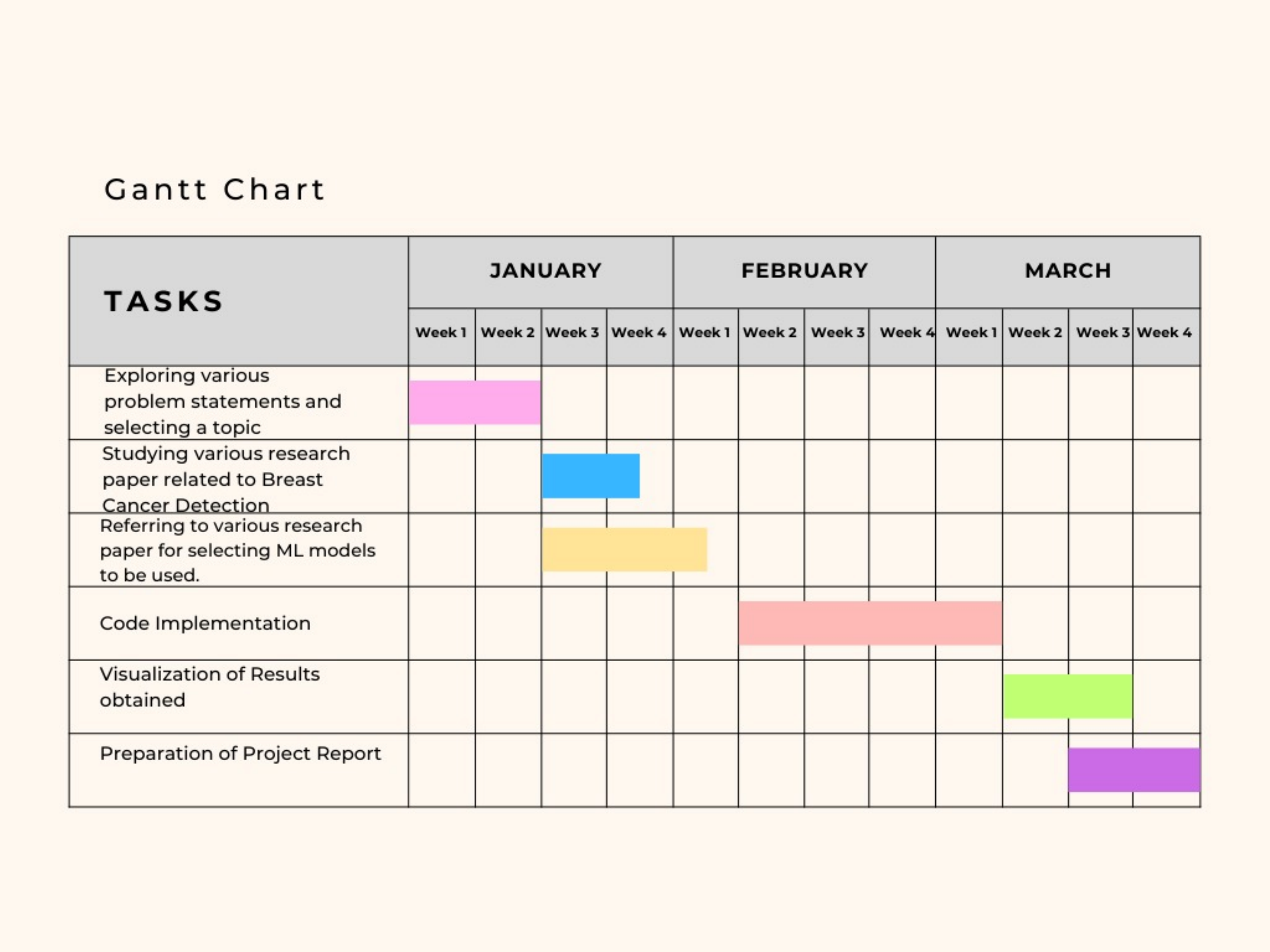


Figure 8 : Flowchart

5.2 Gantt Chart



Gantt Chart

TASKS	JANUARY				FEBRUARY				MARCH			
	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4
Exploring various problem statements and selecting a topic	■	■										
Studying various research paper related to Breast Cancer Detection			■	■								
Referring to various research paper for selecting ML models to be used.			■	■	■							
Code Implementation						■	■	■	■			
Visualization of Results obtained										■	■	
Preparation of Project Report											■	■

Figure 9 : Gantt Chart

Chapter 6: Conclusion and Future Work

6.1 Conclusion

The aim of this project was to develop a machine learning-based system capable of detecting breast cancer using diagnostic features extracted from cell nuclei. The system utilized the Breast Cancer Wisconsin (Diagnostic) dataset and implemented three widely used machine learning models: **Random Forest**, **K-Nearest Neighbors (KNN)**, and **XGBoost**.

After training the models, the system evaluated each classifier using performance metrics such as accuracy, precision, recall, and F1-score. Among all models, **XGBoost achieved the highest accuracy**, demonstrating its effectiveness for this classification task.

This project proves that machine learning techniques can offer a **fast, scalable, and accurate alternative to manual diagnosis**, reducing human dependency and increasing early detection capabilities. By automating the classification process, this system can assist healthcare professionals in making better and faster decisions, especially in resource-constrained environments.

6.2 Future Work

While the current system provides promising results in terms of classification accuracy, there are several opportunities to enhance and extend its functionality:

1. Breast Cancer Stage Prediction

- Extend the model to predict the stage or severity level (Stage I, II, III, IV) of breast cancer, not just benign or malignant classification.
- This could help doctors not only detect cancer but also assess treatment urgency.

2. Integration with Medical Imaging

- Integrate image-based deep learning models such as Convolutional Neural Networks (CNNs) to process mammogram or histopathological images directly.
- This would allow the system to work with raw medical scans in addition to tabular data.

3. Web-based Diagnostic Tool

- Deploy the model in a web or mobile application where users or doctors can upload data or images and receive predictions instantly.
- Include report generation and export features.

4. Multi-class Classification

- Enhance the model to classify other types of tumors or detect multiple cancers.
- Use multi-class ML algorithms for broader diagnostic utility.

5. Real-Time Data Integration

- Integrate the model with Electronic Health Records (EHR) or hospital databases to enable real-time predictions.
- This would streamline the workflow for clinicians.

6. Explainable AI

- Implement explainability tools like SHAP or LIME to visualize which features influenced the prediction.
- This helps build trust among medical practitioners and patients.

7. Larger and More Diverse Datasets

- Train and validate the models on larger and more diverse datasets to improve generalizability and avoid overfitting.
- Collaborate with healthcare institutions to acquire real-world datasets.

Chapter 7: References

- **References**

1. Mahmood, S. A., Al-Battbootti, M. J. H., & Hamadi, S. S. (2025). Breast Cancer Detection Analysis Using Different Machine Learning Techniques: South Iraq Case Study.
2. Al-Nawashi, M. M., Al-Hazaimeh, O. M., & Khazaaleh, M. K. (2024). A New Approach for Breast Cancer Detection Using Machine Learning Techniques.
3. Akash, M. B., Nitheesh, L., Shashank, H. U., & Nagarle, A. (2024). Improved Breast Cancer Detection Using Machine Learning.
4. Bhise, S., Bepari, S., Gadekar, S., Kale, D., & Gaur, A. S. (2023). Breast Cancer Detection Using Machine Learning Techniques.
5. Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast Cancer Detection Using Machine Learning Algorithms. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS).*
6. Al-Imran, M., Akter, S., Mozumder, M. A. S., Bhuiyan, R. J., Rahman, T., Ahmmed, M. J., Hossain Mir, M. N., Hasan, M. A., & Das, A. C. (2024). Evaluating Machine Learning Algorithms for Breast Cancer Detection: A Study on Accuracy and Predictive Performance. *The American Journal of Engineering and Technology.*
7. Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Nagi, M. F. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering.*