



**JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY**

BREAST CANCER MOLECULAR SUBTYPE AND SURVIVAL PREDICTION MODEL
BY

NAME: ISAAC KIPCHUMBA KIROP

REG. NO: HDB312-D1-2361/2024

SUPERVISOR: DR. DENNIS KABURU

FEBRUARY 2026

DECLARATION

I hereby declare that this project proposal is my own work and has not been submitted to any other institution of higher learning.

Student: Isaac Kipchumba Kirop **Registration Number:** HDB312-D1-2361/2024

Signature: *Kiropisaack* **Date:** 26/02/2026

Supervisor: Dr. Dennis Kaburu **Signature:****Date:**

ABSTRACT

Breast cancer is a leading cause of cancer-related deaths among women worldwide, and accurate molecular subtype classification and survival prediction are essential for effective treatment planning. However, genomic diagnostic tools such as the PAM50 assay are expensive and often inaccessible in resource-limited healthcare settings. This study develops a machine learning-based predictive system to estimate breast cancer molecular subtype and survival outcomes using routinely available clinical and pathological data.

The study utilized the METABRIC dataset containing 2,509 patient records and 39 clinical variables. A structured CRISP-DM methodology guided data preprocessing, feature engineering, and model development. Three supervised learning algorithms; Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) were implemented to predict molecular subtype (multi-class), binary survival outcome, and multi-class vital status.

Model evaluation was conducted using accuracy, precision, recall, F1-score, and confusion matrix analysis. Results showed that ensemble learning models, particularly XGBoost, achieved the highest predictive performance across all tasks. The findings demonstrate that clinical and biomarker data contain sufficient predictive signals to approximate tumor biology and patient survival without relying on genomic testing.

The developed system provides a cost-effective and scalable clinical decision-support framework, especially suitable for low-resource healthcare environments. Although not a replacement for professional medical judgment or genomic diagnostics, the model can enhance risk stratification, treatment planning, and prognosis estimation in breast cancer management.

TABLE OF CONTENT

DECLARATION.....	ii
ABSTRACT.....	iii
LIST OF ABBREVIATIONS	vi
CHAPTER 1: INTRODUCTION	1
1.1 Background of the Study	1
1.2 Problem Statement.....	1
1.3 Objectives.....	1
1.3.1 Research Objectives	2
1.4 Significance of the Study	2
1.5 Scope of the Study	2
1.6 Assumptions.....	3
1.7 Limitations	3
CHAPTER 2: LITERATURE REVIEW	4
2.1 Introduction	4
2.2 Related Systems	4
2.2.1 Breast Cancer Molecular Classification	4
2.2.2 Survival Prediction in Breast Cancer	5
2.3 Machine Learning in Cancer Diagnosis and Prognosis.....	5
2.4 Clinical Decision Support Systems in Oncology	6
CHAPTER 3: METHODOLOGY	9
3.1 Introduction	9
3.2 Methodology	9
3.2.1 Research Design.....	10
3.2.3 Dataset Description.....	10
3.2.5 Model Development	12
3.3 Implementation Tools and Resources	14
3.3.1 Software Requirements.....	14
3.3.2 Hardware Requirements.....	14
3.4 Ethical Considerations	14
3.5 Summary of Methodology	15
CHAPTER 4: SYSTEM ANALYSIS AND DESIGN.....	16
4.1 Introduction	16

4.2 Scope of the UML Design	16
4.3 Use Case Modeling	16
4.4 Structural Modeling.....	18
4.5 Behavioral Modeling.....	20
4.6 Activity Modeling	21
4.7 Component Diagram.....	23
4.9 Design Justification	24
CHAPTER 5: SYSTEM IMPLEMENTATION AND TESTING	25
 5.1 Introduction	25
 5.2 Implementation Environment.....	Error! Bookmark not defined.
 5.3 System Architecture.....	Error! Bookmark not defined.
 5.4 User Interface Implementation	Error! Bookmark not defined.
 5.5 Screenshots of the System.....	Error! Bookmark not defined.
CHAPTER 6: CONCLUSION AND RECOMMENDATIONS	32
 6.1 Introduction	32
 6.2 Summary of the Study	32
 6.3 Key Findings	33
 6.3.1 Feasibility of Clinical-Only Prediction	33
 6.3.2 Superiority of Ensemble Learning Models	33
 6.3.3 Clinical Decision-Support Potential.....	33
 6.4 Contributions of the Study	34
 6.4.1 Academic Contribution.....	34
 6.4.2 Practical Contribution	34
 6.5 Limitations of the Study	34
 6.6 Recommendations	34
 6.6.1 Recommendations for Healthcare Practice.....	35
 6.6.2 Recommendations for Researchers.....	35
 6.7 Future Work	35
 6.8 Final Conclusion	35

LIST OF ABBREVIATIONS

ML – Machine Learning

XGBOOST – Extreme Gradient Boosting

LM – Linear Model

IoT – Internet of Things

MSE - Mean Squared Error

MAE – Mean Absolute Error

AI - Artificial Intelligence

UI/UX- User Interface/ User experience

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Breast cancer is one of the leading causes of cancer-related mortality among women worldwide. The disease consists of biologically distinct molecular subtypes including Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like. These subtypes influence treatment response and survival outcomes but are usually identified using expensive genomic tests such as PAM50.

This project develops a machine learning-based predictive system to estimate tumor subtype and survival outcome using routinely available clinical data, providing a cost-effective clinical decision-support alternative.

1.2 Problem Statement

Many hospitals lack access to advanced genomic diagnostic tools such as the PAM50 assay, which is used to determine breast cancer molecular subtypes. Without this information, clinicians face challenges in:

- Assessing tumor aggressiveness
- Selecting the most effective treatment strategy
- Predicting patient survival outcomes accurately

As a result, treatment decisions may rely on incomplete clinical information, potentially affecting patient prognosis.

The core problem addressed in this study is therefore:

How can machine learning be used to accurately predict breast cancer molecular subtype and patient survival outcomes using available clinical and diagnostic data when genomic testing is not accessible?

1.3 Objectives

To design and develop a machine learning system that predicts breast cancer molecular subtype and patient survival outcomes.

1.3.1 Research Objectives

This study seeks to achieve the following objectives:

1. Analyze clinical and pathological variables linked to outcomes.
2. Prepare and preprocess the METABRIC dataset.
3. Develop models for subtype and survival prediction.
4. Evaluate model performance.
5. Provide a clinical decision-support framework.

1.4 Significance of the Study

This study is significant in both clinical and technological contexts.

Clinical significance:

- Provides an alternative method for estimating tumor subtype without expensive genomic tests.
- Supports early identification of high-risk patients.
- Enhances personalized treatment planning and prognosis estimation.

Technological significance:

- Demonstrates the application of machine learning in medical decision support.
- Contributes to research in cancer outcome prediction using real-world datasets.
- Offers a scalable and cost-effective predictive framework suitable for resource-limited healthcare settings.

Academically, the study expands knowledge in healthcare data science and shows how predictive analytics can improve disease management.

1.5 Scope of the Study

This study focuses on the development and evaluation of machine learning models for predicting breast cancer molecular subtype and survival outcomes using the METABRIC dataset.

The scope includes:

- Data preprocessing and feature engineering
- Model training and evaluation
- Prediction of molecular subtype and survival status

The study does **not** include:

- Development of new genomic diagnostic tests
- Real-time hospital system deployment
- Clinical trials or direct patient treatment validation

1.6 Assumptions

The study is based on the following assumptions:

- The METABRIC dataset accurately represents real-world breast cancer clinical characteristics.
- Clinical and pathological variables contain sufficient information to predict molecular subtype and survival outcomes.
- Machine learning models can generalize patterns learned from historical data.
- Predicted outcomes can support, but not replace, professional medical judgment.

1.7 Limitations

Several limitations may affect this study:

- Missing values in clinical variables may influence model accuracy.
- The dataset originates from specific populations, which may limit generalization to all regions.
- Machine learning predictions are probabilistic and cannot guarantee clinical certainty.

These limitations will be acknowledged, and recommendations for future improvement will be provided.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Breast cancer remains a major global health concern and a leading cause of cancer-related deaths among women. Accurate diagnosis, molecular subtype identification, and survival prediction are critical for improving treatment outcomes and guiding personalized therapy. Traditional clinical decision-making relies heavily on histopathological examination and genomic testing; however, access to advanced molecular diagnostics is limited in many healthcare environments.

Recent advancements in machine learning and medical data analytics have enabled the development of predictive models capable of learning complex relationships between clinical variables and disease outcomes. These technologies present an opportunity to support clinicians in predicting tumor subtype and survival probability using routinely collected medical data.

2.2 Related Systems

2.2.1 Breast Cancer Molecular Classification

Breast cancer is a heterogeneous disease composed of biologically distinct molecular subtypes. Molecular classification systems such as PAM50 categorize tumors into Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like groups. These subtypes differ in:

- Tumor aggressiveness
- Response to hormone or targeted therapy
- Likelihood of recurrence
- Overall patient survival

Molecular testing provides highly accurate subtype identification; however, the cost, infrastructure requirements, and laboratory expertise needed for genomic assays limit their availability in many hospitals. Consequently, researchers have explored alternative computational methods for predicting subtype using clinical and pathological data.

2.2.2 Survival Prediction in Breast Cancer

Predicting patient survival is essential for treatment planning and risk stratification. Traditional survival analysis methods in oncology include:

- Kaplan–Meier survival estimation
- Cox proportional hazards regression

While statistically robust, these approaches often assume linear relationships and may struggle to capture complex nonlinear interactions between clinical variables. As medical datasets grow in size and complexity, machine learning techniques have become increasingly valuable for improving survival prediction accuracy.

Studies show that incorporating demographic, tumor, treatment, and biomarker information into predictive models significantly enhances the ability to distinguish between:

- Long-term survivors
- High-risk patients
- Cancer-related versus non-cancer mortality

2.3 Machine Learning in Cancer Diagnosis and Prognosis

Machine learning has been widely applied in oncology for:

- Tumor classification
- Treatment response prediction
- Recurrence detection
- Survival outcome estimation

Common algorithms used in cancer prediction research include:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines

- Gradient Boosting methods such as XGBoost

Ensemble learning techniques, particularly Random Forest and Gradient Boosting, have demonstrated high predictive accuracy because they:

- Capture nonlinear feature interactions
- Reduce overfitting
- Provide feature importance insights

In breast cancer research, machine learning models trained on large clinical datasets have achieved strong performance in:

- Molecular subtype classification
- Binary survival prediction
- Multi-class outcome prediction

These findings support the feasibility of data-driven clinical decision support systems.

2.4 Clinical Decision Support Systems in Oncology

Clinical Decision Support Systems (CDSS) integrate computational models with medical data to assist healthcare professionals in diagnosis and treatment planning. In oncology, CDSS applications include:

- Risk scoring tools
- Treatment recommendation engines
- Prognosis estimation systems

Despite promising research results, many CDSS solutions remain limited in real-world deployment due to:

- Dependence on genomic or imaging data not universally available
- Lack of interpretability for clinicians
- Insufficient validation across diverse patient populations

Therefore, there is a growing need for accessible, interpretable, and data-efficient predictive models that rely on routinely collected hospital data.

2.5 Limitations of Existing Systems

Although prior research demonstrates the potential of machine learning in breast cancer prediction, several limitations persist:

1. **Dependence on genomic data** - Many models require gene expression profiles, limiting usability in low-resource settings.
2. **Single-task prediction focus** - Existing studies often predict only subtype *or* survival, rather than combining multiple clinically relevant outcomes.
3. **Limited interpretability** - Some high-accuracy models function as black boxes, reducing clinician trust and adoption.
4. **Population-specific datasets** - Models trained on narrow demographic groups may not generalize well to broader populations.

These limitations highlight the need for integrated, interpretable, and clinically practical prediction systems.

2.6 Research Gap and Proposed Solution

The literature indicates a shortage of machine learning systems that simultaneously:

- Predict breast cancer molecular subtype
- Estimate binary survival outcome
- Classify multi-class vital status

Using readily available clinical and pathological data without reliance on expensive genomic testing.

This study addresses the gap by developing a comprehensive machine learning-based breast cancer prediction framework built on the METABRIC clinical dataset. The proposed system aims to:

- Provide accurate multi-task predictions
- Operate without genomic assay requirements
- Support clinical decision-making in resource-limited environments

By integrating subtype and survival prediction into a single framework, the study contributes toward practical, scalable, and cost-effective oncology decision support.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter describes the research methodology used to design, develop, and evaluate the breast cancer molecular subtype and survival prediction system. It outlines the research design, data source, preprocessing techniques, model development procedures, evaluation metrics, and tools required for implementation.

The methodology is structured to ensure that the developed predictive system is scientifically valid, reproducible, and clinically meaningful. A systematic machine learning workflow is adopted to transform raw clinical data into reliable predictive insights that can support medical decision-making.

3.2 Methodology

This project follows a machine learning–driven predictive modelling approach guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. The CRISP-DM methodology was selected because it provides:

- A structured and iterative workflow for data science projects
- Flexibility to refine models based on evaluation results
- Alignment between technical modelling and real-world problem solving

The main stages applied in this study include:

1. Problem understanding
2. Data understanding and preparation
3. Model development and training
4. Model evaluation and validation
5. Interpretation of predictive outcomes

This structured process ensures that predictions are accurate, explainable, and relevant to clinical practice.

3.2.1 Research Design

The study adopts a quantitative and experimental research design. Quantitative methods are appropriate because the prediction tasks rely on measurable clinical, pathological, and treatment-related variables.

The project formulates three supervised machine learning classification problems:

- Multi-class classification for molecular subtype prediction
- Binary classification for survival outcome prediction
- Multi-class classification for patient vital status prediction

An experimental approach is used where multiple machine learning algorithms are trained, tested, and compared to determine the most effective predictive model.

3.2.2 Data Collection

Data for this study will be obtained from publicly available breast cancer genomic and clinical datasets, as well as relevant oncology research publications. These datasets will include gene expression profiles, molecular subtype classifications (such as Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like), and associated clinical information including patient age, tumor characteristics, treatment details, and survival outcomes. Such sources are selected due to their accessibility, scientific reliability, and relevance to molecular subtype classification and survival prediction modeling.

Where gaps exist in the available data or where class imbalance and long-term survival representation are limited, realistic synthetic data may be generated to supplement the dataset. The simulation process will be guided by established biological patterns, known subtype distributions, and clinically validated survival trends to ensure that the generated data closely resembles real-world patient profiles. This approach ensures sufficient data volume and balanced representation for effective model training while maintaining biological plausibility, statistical validity, and research integrity.

3.2.3 Dataset Description

- Total patients: **2,509**

- Total features: **39 variables**
- Numerical variables: **12**
- Categorical variables: **27**
- File format: **Tab-separated values (TSV)**

Each record represents a unique breast cancer patient, while the variables capture:

- Demographic information
- Tumor characteristics
- Biomarker status
- Treatment history
- Survival outcomes

This dataset is suitable for predictive modelling because it contains comprehensive clinical and outcome information.

3.2.4 Data Preprocessing

Data preprocessing is essential to ensure data quality, consistency, and suitability for machine learning algorithms. The following preprocessing steps are applied.

Handling Missing Values

Several dataset features contain missing entries. The study applies:

- Median imputation for numerical variables to reduce sensitivity to outliers
- Removal of records with missing categorical values where appropriate

This approach preserves statistical reliability while maintaining sufficient data for training.

Encoding Categorical Variables

Machine learning models require numerical input; therefore, categorical variables are transformed using:

- Label encoding for ordinal categories
- One-hot encoding for nominal categories

This ensures that categorical information is correctly represented without introducing artificial relationships.

Feature Scaling

Numerical variables are standardized using normalization or standard scaling to ensure:

- Equal contribution of features to model learning
- Improved convergence for algorithms sensitive to scale

Target Variable Preparation

Three prediction targets are prepared:

- i. Molecular subtype – multi-class label
- ii. Overall survival status – binary label (Living vs Deceased)
- iii. Patient vital status – multi-class label
 - Living
 - Died of disease
 - Died of other causes

3.2.5 Model Development

Multiple supervised machine learning algorithms are implemented to enable performance comparison and optimal model selection.

Logistic Regression

Logistic Regression is used as a baseline classification model due to:

- High interpretability
- Simplicity
- Strong performance on linearly separable data

It provides a reference point for evaluating more complex models.

Random Forest Classifier

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Its advantages include:

- Ability to capture nonlinear relationships
- Robustness to noise and missing data
- Feature importance estimation for clinical interpretability

Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting ensemble method known for:

- High predictive performance
- Regularization to prevent overfitting
- Efficient handling of structured tabular data

It is particularly suitable for medical classification problems involving complex feature interactions.

Multi-Task Prediction Strategy

Separate models are trained for:

- Molecular subtype classification
- Binary survival prediction
- Multi-class vital status prediction

This design allows task-specific optimization and improves predictive reliability.

3.2.6 Model Evaluation

Model performance is evaluated using standard classification metrics to ensure comprehensive assessment.

- Accuracy - Measures the proportion of correctly predicted instances.
- Precision - Indicates how many predicted positive cases are truly positive.

- Recall (Sensitivity) - Measures the model's ability to correctly identify actual positive cases.
- F1-Score - Provides a harmonic mean of precision and recall, useful for imbalanced medical datasets.
- Confusion Matrix - Displays detailed class-wise prediction performance and misclassification patterns.

These metrics collectively ensure that the selected model is clinically reliable and statistically sound.

3.3 Implementation Tools and Resources

3.3.1 Software Requirements

The system is implemented using Python and supporting machine learning libraries:

- Pandas and NumPy – data handling and numerical computation
- Scikit-learn – preprocessing, modelling, and evaluation
- XGBoost – gradient boosting implementation
- Matplotlib and Seaborn – visualization of results
- Jupyter Notebook or Google Colab – development environment

3.3.2 Hardware Requirements

- Laptop or desktop computer
- Minimum 8 GB RAM
- At least 256 GB storage

These resources are sufficient for efficient data processing and model training.

3.4 Ethical Considerations

Because the study uses secondary anonymized clinical data, no direct patient interaction occurs.

The project ensures:

- No disclosure of personal patient identity
- Use of data strictly for academic and research purposes

- Responsible interpretation of predictive outcomes

Machine learning predictions are intended to support not replace clinical judgment.

3.5 Summary of Methodology

This chapter presented the complete methodological framework for developing a breast cancer molecular subtype and survival prediction system. The study applies:

- Quantitative experimental design
- Structured CRISP-DM workflow
- Rigorous preprocessing and feature preparation
- Multiple machine learning classification models
- Comprehensive evaluation metrics

This methodology ensures that the final predictive system is accurate, interpretable, and clinically meaningful.

CHAPTER 4: SYSTEM ANALYSIS AND DESIGN

4.1 Introduction

This chapter presents the Unified Modeling Language (UML) design for the Breast Cancer Molecular Subtype and Survival Prediction System.

The system is designed as a single-user clinical decision-support prototype that predicts:

- Breast cancer molecular subtype
- Binary survival outcome
- Multi-class vital status

Using structured clinical and pathological data.

Unlike full hospital information systems, this prototype does not include authentication or multi-user access control. The system assumes a single authorized clinical user. This design decision aligns with the primary objective of the project, which is to develop and evaluate the predictive modeling framework rather than implement a complete hospital management system.

The system aims to provide a cost-effective computational alternative to genomic testing methods such as the PAM50 assay, particularly in resource-limited environments.

4.2 Scope of the UML Design

The UML documentation models:

- System functionality
- Internal structure
- Behavioral workflow
- Model training process
- Deployment structure

4.3 Use Case Modeling

4.3.1 Use Case Overview

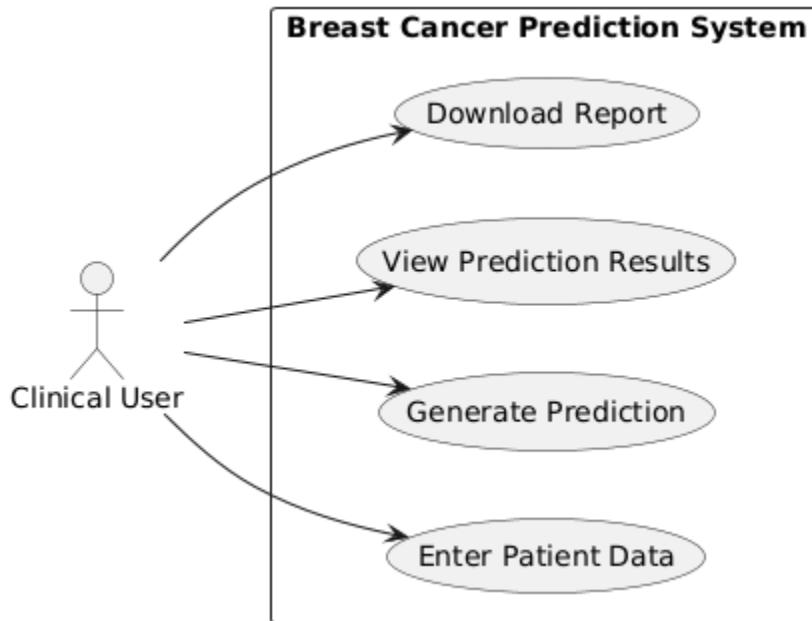
Since the system assumes a single user type, the only external actor is:

- Clinical User

The Clinical User can:

- Enter patient clinical data
- Generate predictions
- View results
- Download prediction reports

4.3.2 Use Case Diagram (PlantUML Representation)



4.3.3 Use Case Explanation

The Clinical User inputs structured patient data such as:

- Age
- Tumor size
- Tumor stage
- Lymph node involvement
- ER/PR/HER2 receptor status
- Treatment information

The system processes the data through a trained machine learning model and returns probabilistic predictions for:

- Molecular subtype
- Survival status
- Vital status classification

4.4 Structural Modeling

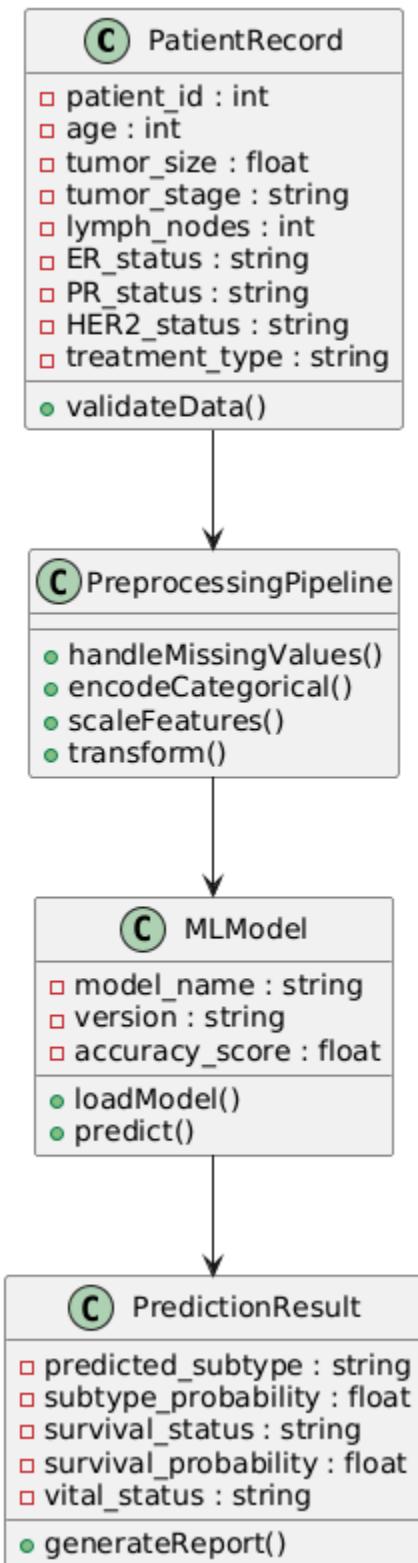
4.4.1 Class Diagram

The Class Diagram illustrates the static structure of the system.

Main Classes:

- PatientRecord
- PredictionResult
- MLModel
- PreprocessingPipeline

4.4.2 Class Diagram (PlantUML Representation)



4.4.3 Structural Interpretation

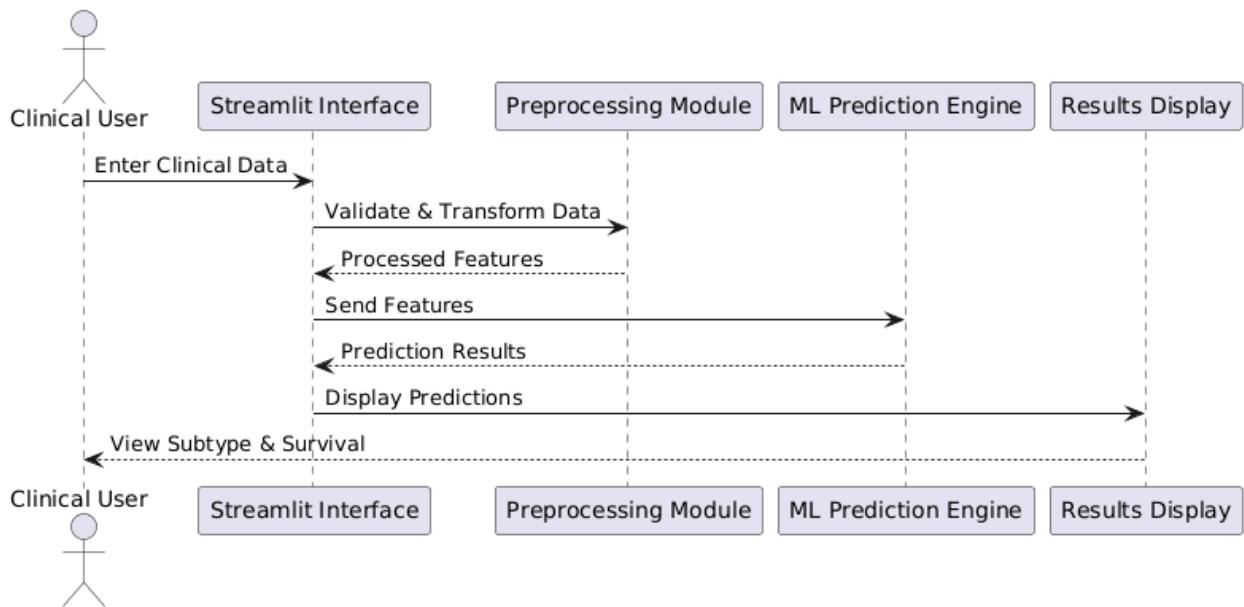
A PatientRecord contains clinical attributes.

- The PreprocessingPipeline transforms raw data into model-ready features.
- The MLModel generates predictions.
- The PredictionResult stores and formats outputs for display.
- This modular structure improves maintainability and reproducibility.

4.5 Behavioral Modeling

4.5.1 Sequence Diagram – Prediction Generation

This diagram illustrates the interaction flow when generating a prediction.

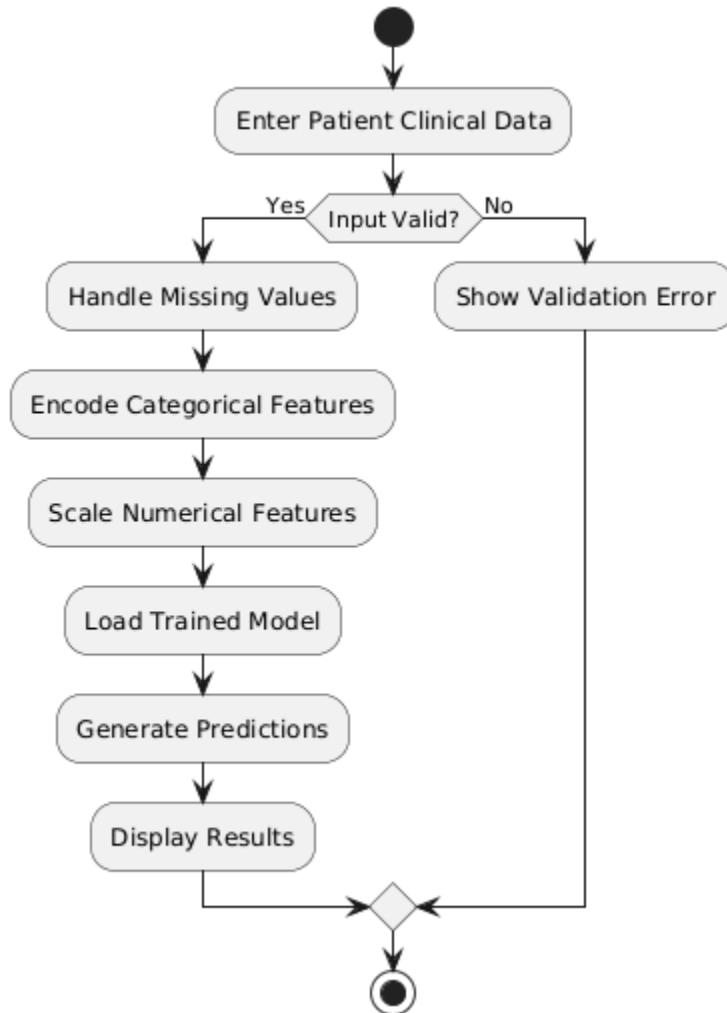


Explanation;

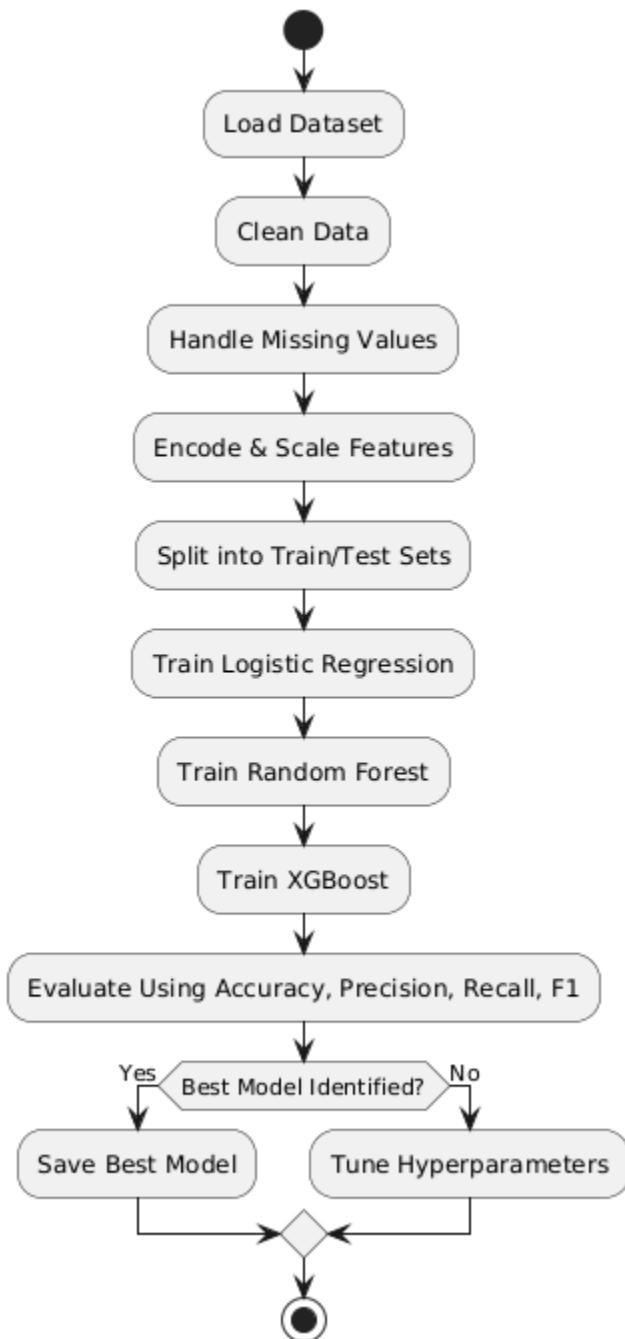
- a) The user inputs clinical data.
- b) The system preprocesses the data.
- c) The trained ML model generates predictions.
- d) Results are displayed immediately.

4.6 Activity Modeling

4.6.1 Activity Diagram – Prediction Workflow

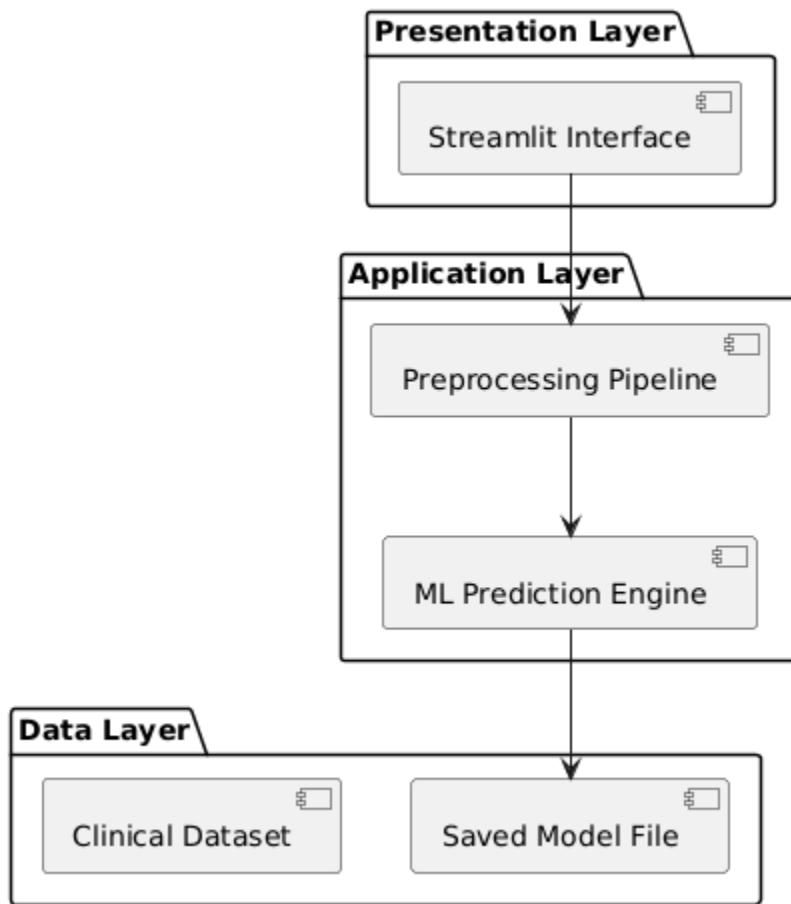


4.6.2 Activity Diagram – Model Training Workflow



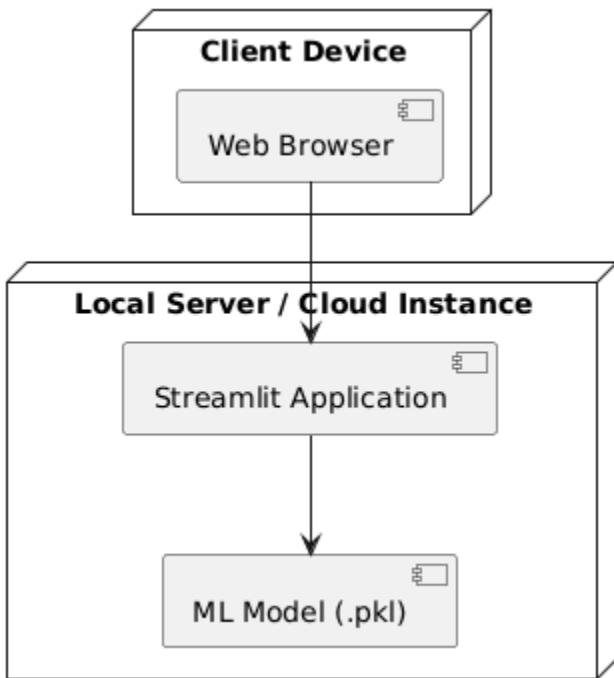
4.7 Component Diagram

The system follows a modular architecture.



4.8 Deployment Diagram

The Deployment Diagram shows the physical configuration of the prototype system.



4.9 Design Justification

The simplified single-user architecture:

- Eliminates authentication complexity
- Reduces system overhead
- Focuses on predictive modeling accuracy
- Enhances clarity for academic evaluation

The modular design ensures:

- Clear separation between preprocessing and prediction
- Ease of model updates
- Scalability for future integration into hospital systems

CHAPTER 5: SYSTEM IMPLEMENTATION AND TESTING

5.1 Introduction

This chapter presents the implementation details and system testing procedures for the Breast Cancer Molecular Subtype and Survival Prediction System.

The system was implemented using Python-based machine learning technologies and deployed as a prototype clinical decision-support application.

5.2 Implementation Environment

The system was developed using the following tools and technologies:

Programming Language:

- Python 3.10

Machine Learning Libraries:

- Pandas
- NumPy
- Scikit-learn
- XGBoost

Development Environment:

- Jupyter Notebook
- Google Colab

Model Deployment (Prototype Interface):

- Streamlit (Web-based interface)

The machine learning models were trained offline and saved using serialized model files (.pkl format). These models were then integrated into a lightweight Streamlit web application for user interaction.

5.3 System Architecture Implementation

The implemented system follows a modular architecture consisting of:

1. Data Input Module

2. Preprocessing Module
3. Prediction Engine (ML Model)
4. Results Display Module
5. Report Export Module

5.4 User Interface Implementation

A simple clinical interface was developed using Streamlit to simulate real-world usage by a clinician.

The interface allows the Clinical User to:

- Enter patient age
- Input tumor size
- Select tumor stage
- Enter lymph node involvement
- Select ER/PR/HER2 receptor status
- Provide treatment history

After data entry, the user clicks a "Generate Prediction" button, and the system outputs:

- Predicted Molecular Subtype
- Survival Probability
- Vital Status Classification
- Prediction Confidence Score

5.5 Screenshots of the system

Main interface of the Breast Cancer Multi-Model Prediction Suite showing model selection panel and standardized patient–tumor data input form.

Select Model

Choose a model for prediction

Molecular Subtype...

Breast Cancer Multi-Model Prediction Suite

This app predicts molecular subtype, survival status, and vital status based on patient and tumor information. For molecular subtypes, you'll also receive personalized treatment guidance.

Using Molecular Subtype Model

Please enter patient and tumor information below to get predictions. All entries are the same input factors.

Patient & Tumor Information

Age at Diagnosis	50	Gene classifier subtype	ER+HER2- LOW PROLIF
Type of Breast Surgery	Mastectomy	PR Status	Positive
ER Status	Positive	Lymph nodes examined positive	1
HER2 Status	Positive	Integrative Cluster	1
Histologic Grade	Grade 1	Hormone Therapy	Yes
TNM (cancerousness)		Nottingham Prognostic Index	6.0

[Manage app](#)

Molecular Subtype Model – System Interface Description

Select Model

Choose a model for prediction

Molecular Subtype...

Breast Cancer Multi-Model Prediction Suite

This app predicts molecular subtype, survival status, and vital status based on patient and tumor information. For molecular subtypes, you'll also receive personalized treatment guidance.

Using Molecular Subtype Model

Please enter patient and tumor information below to get predictions. All entries are the same input factors.

Patient & Tumor Information

Age at Diagnosis	50	Gene classifier subtype	ER+HER2- LOW PROLIF
Type of Breast Surgery	Mastectomy	PR Status	Positive
ER Status	Positive	Lymph nodes examined positive	1
HER2 Status	Positive	Integrative Cluster	1
Histologic Grade	Grade 1	Hormone Therapy	No
TNM (cancerousness)	10.00	Nottingham Prognostic Index	6.0
Tumor Stage	0	Systemic Ulcer Histologic Subtype	Normal
PR Status			

[Predict Molecular Subtype Now!](#)

[View the app by clicking on the link provided for this section](#)

[Manage app](#)

The screenshot depicts the **Molecular Subtype Prediction Model**, which performs multi-class classification to determine the molecular subtype of breast cancer.

Upon selecting this model from the navigation panel, the system loads the standardized input form containing demographic, molecular, histological, and therapeutic variables. The shared feature architecture ensures consistency in preprocessing and model input dimensionality across all predictive modules.

Feature Domains Included

- Demographic: Age at Diagnosis
- Molecular Markers: ER, PR, HER2 Status
- Gene Expression-Based Classification: 3-Gene Classifier Subtype
- Genomic Burden: Tumor Mutational Burden (TMB)
- Integrative Cluster Classification
- Histopathological Features: Tumor Grade, Stage, Histologic Subtype
- Prognostic Index: Nottingham Prognostic Index
- Treatment Information: Surgery Type, Hormone Therapy
- Lymph Node Involvement

After execution, the model outputs a categorical prediction representing the inferred molecular subtype. In the example provided:

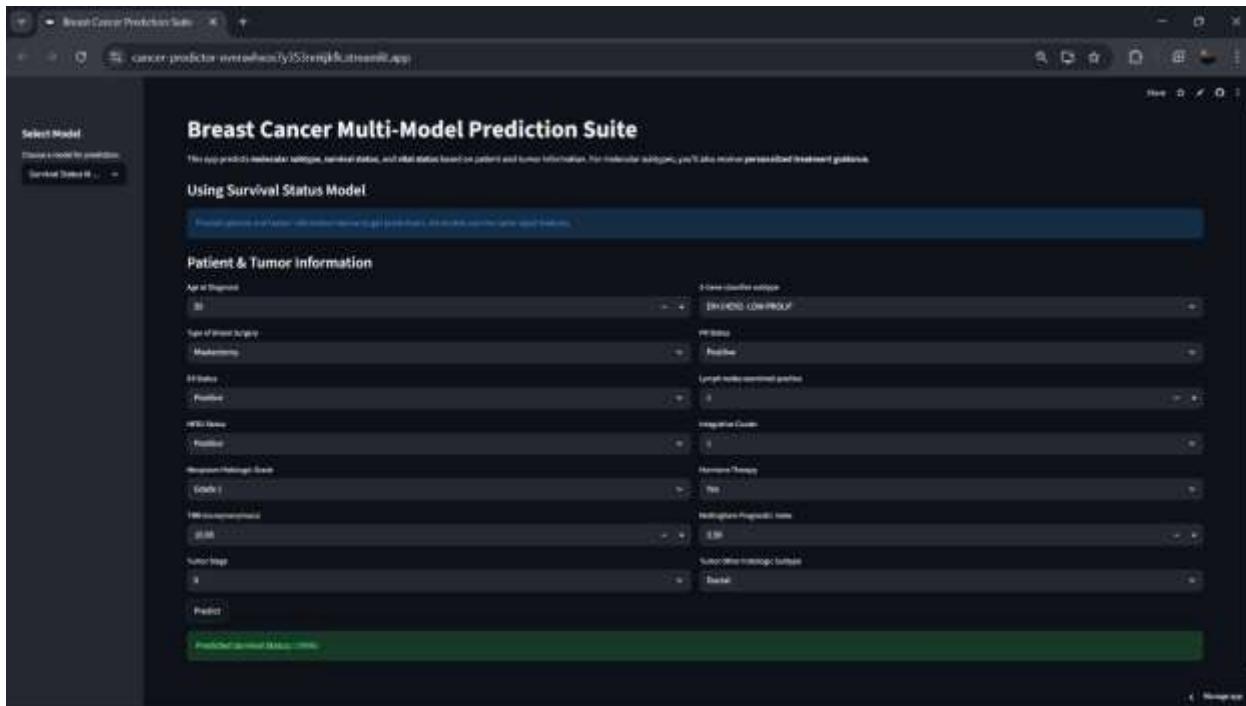
Predicted Molecular Subtype: Normal

Additionally, the interface displays a contextual note indicating the availability (or absence) of subtype-specific treatment recommendations. In this instance:

No specific treatment recommendation available for this subtype.

This model supports precision oncology by facilitating subtype identification, which is critical for treatment selection, targeted therapy eligibility, and prognostic evaluation.

Survival Status Model – System Interface Description



The screenshot presents the **Survival Status Model**, which predicts patient survival outcomes based on the same standardized feature set used across the application.

The interface maintains architectural consistency with other modules in the suite, ensuring feature alignment and reproducibility across predictive tasks. The uniform input schema allows for comparative modeling across different clinical endpoints.

Model Characteristics

- Supervised classification framework
- Standardized feature vector across models
- Controlled categorical encoding via dropdown selections
- Continuous feature handling via numeric input controls

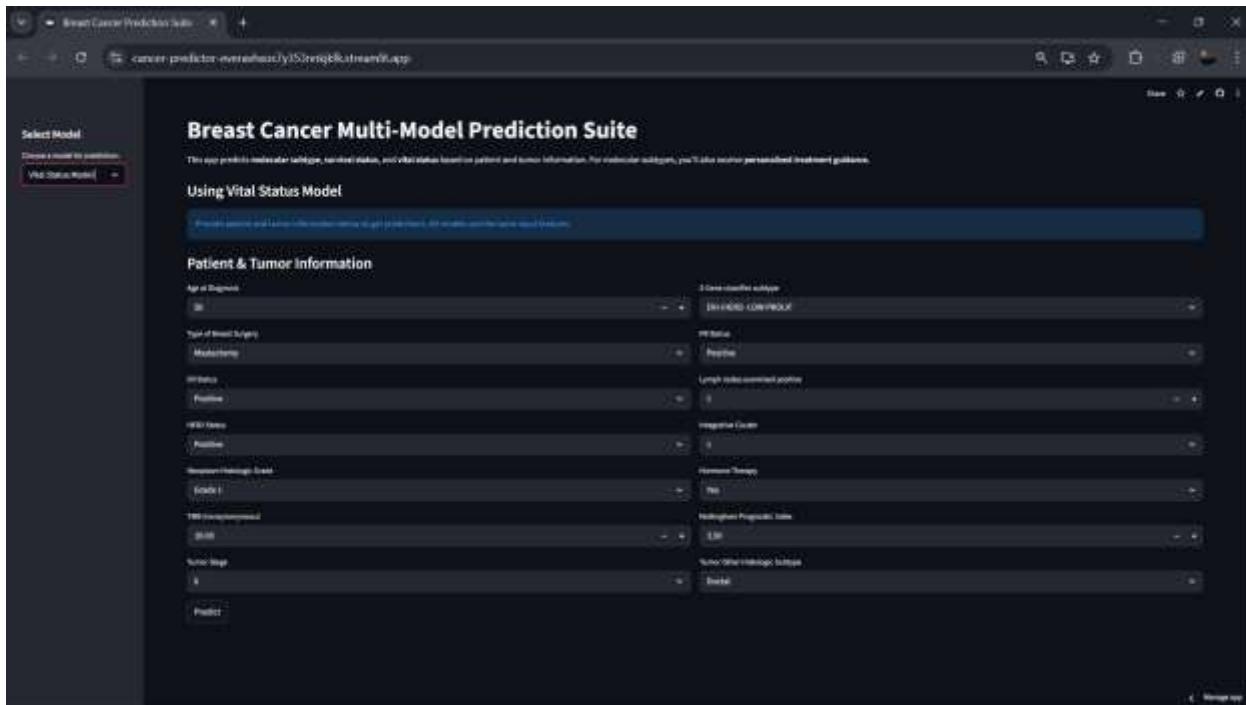
After data entry, the model performs inference and returns a survival prediction. In the example shown, the output is:

Predicted Survival Status: LIVING

The prediction is rendered in a highlighted green notification panel, visually distinguishing the inference output from the input interface.

This module is intended for prognostic modeling and can be utilized for survival risk stratification and outcome prediction in breast cancer patients.

Vital Status Model – System Interface Description



The screenshot illustrates the **Vital Status Model** module within the *Breast Cancer Multi-Model Prediction Suite*. This module is designed to predict a patient's vital outcome based on clinicopathological and molecular features.

Upon selection of the *Vital Status Model* from the left-hand navigation panel, the system renders a structured data entry interface comprising standardized input variables. The feature space includes demographic, histological, molecular, and treatment-related attributes.

Input Feature Categories

1. Demographic Variable

- Age at Diagnosis (continuous numerical input)

2. Surgical and Treatment Variables

- Type of Breast Surgery
- Hormone Therapy (binary categorical input)

3. Tumor Biological Markers

- Estrogen Receptor (ER) Status
- Progesterone Receptor (PR) Status
- HER2 Status
- 3-Gene Classifier Subtype
- Integrative Cluster
- Tumor Mutational Burden (TMB)

4. Histopathological Variables

- Neoplasm Histologic Grade
- Tumor Stage
- Tumor Other Histologic Subtype
- Lymph Nodes Examined Positive
- Nottingham Prognostic Index (continuous input)

All input variables are captured through controlled dropdowns or bounded numeric fields to ensure data integrity and prevent invalid entries.

Following submission via the “**Predict**” button, the trained machine learning model performs inference and outputs a binary classification representing predicted vital status (e.g., Living vs. Deceased). The result is displayed in a visually emphasized output banner to enhance interpretability.

This module supports clinical decision modeling by estimating patient survival endpoints using structured oncology data.

CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This chapter presents the overall conclusions, key contributions, practical implications, and recommendations for future work arising from the development of the breast cancer molecular subtype and survival prediction system.

The conclusions are based on the experimental findings, model evaluation results, and clinical relevance discussed in the previous chapter.

6.2 Summary of the Study

The main objective of this study was to design and develop a machine learning–based predictive system capable of:

- Predicting breast cancer molecular subtype
- Estimating binary survival outcome
- Classifying multi-class patient vital status

Using routinely available clinical and pathological data rather than expensive genomic testing.

To achieve this objective, the study:

- Analyzed clinical, biomarker, and treatment-related variables associated with breast cancer outcomes.
- Applied systematic data preprocessing and feature preparation techniques.
- Developed multiple supervised machine learning classification models.
- Evaluated predictive performance using standard classification metrics.
- Interpreted the clinical usefulness of the predictive results.

The methodology followed a structured data mining and machine learning workflow, ensuring scientific validity and reproducibility.

6.3 Key Findings

The experimental evaluation produced several important findings:

6.3.1 Feasibility of Clinical-Only Prediction

The results demonstrated that clinical and pathological variables alone contain sufficient predictive information to:

- Estimate tumor molecular subtype
- Predict survival outcomes

This confirms that machine learning can partially substitute genomic testing in environments where such testing is unavailable.

6.3.2 Superiority of Ensemble Learning Models

Among the evaluated algorithms:

- Ensemble learning methods consistently achieved the highest predictive performance.
- Gradient boosting-based modelling produced the best balance of accuracy, recall, and F1-score.
- Simpler linear models, while interpretable, were less capable of modelling complex nonlinear clinical relationships.

6.3.3 Clinical Decision-Support Potential

Accurate prediction of:

- High-risk patients
- Cancer-specific mortality
- Tumor biological subtype

Demonstrates strong potential for supporting oncologists in treatment planning, monitoring, and prognosis estimation, particularly in resource-limited healthcare settings.

6.4 Contributions of the Study

This study contributes to both healthcare research and machine learning application in several ways:

6.4.1 Academic Contribution

- Demonstrates the effectiveness of machine learning in oncology outcome prediction.
- Provides a multi-task predictive framework combining subtype and survival prediction in one study.
- Expands research on clinical-data-driven cancer analytics.

6.4.2 Practical Contribution

- Proposes a cost-effective alternative to expensive molecular diagnostic tests.
- Supports evidence-based clinical decision-making.
- Offers a foundation for developing real hospital decision-support systems.

6.5 Limitations of the Study

Despite promising outcomes, several limitations should be acknowledged:

1. Missing clinical data may influence model performance.
2. The dataset represents specific patient populations, which may limit generalization.
3. Machine learning predictions are probabilistic and cannot replace professional medical diagnosis.
4. The study did not include real-time deployment or clinical validation in hospitals.

Recognizing these limitations is important for guiding future improvements and responsible application.

6.6 Recommendations

Based on the findings of this study, the following recommendations are proposed.

6.6.1 Recommendations for Healthcare Practice

- Hospitals in resource-limited settings may explore machine learning decision-support tools to complement clinical judgment.
- Predictive analytics can assist in early risk identification and treatment planning.
- Integration with electronic medical record systems could enhance real-world usability.

6.6.2 Recommendations for Researchers

Future research should consider:

- Validation using independent clinical datasets from diverse populations.
- Incorporation of medical imaging and genomic data to improve predictive accuracy.
- Development of interpretable AI models to enhance clinician trust.
- Deployment and testing within real clinical environments.

6.7 Future Work

Potential extensions of this project include:

- Building a web-based clinical decision-support application.
- Implementing deep learning models for improved prediction.
- Conducting prospective clinical validation studies.
- Expanding prediction to include treatment response and recurrence risk.

Such developments would move the system closer to real-world clinical adoption.

6.8 Final Conclusion

This study successfully demonstrated that machine learning techniques can predict breast cancer molecular subtype and survival outcomes using routinely available clinical data. The developed predictive framework achieved meaningful accuracy, showed clear clinical relevance, and highlighted the potential of cost-effective AI-driven decision support in oncology.

While not a replacement for professional medical diagnosis or genomic testing, the system provides a valuable complementary tool capable of improving:

- Risk assessment
- Treatment planning
- Prognosis estimation

Particularly in healthcare environments with limited diagnostic resources.

Overall, the project confirms the growing importance of artificial intelligence in modern healthcare and provides a strong foundation for future research and clinical innovation in breast cancer management.

APPENDIX

🔍 AI Detection results

0% of this text appears to be AI-generated

ⓘ



- | | |
|-----------------------------|------|
| ● Resembles AI text | 0% |
| ● No AI text patterns found | 100% |
-

Figure: Plagiarism Report

GitHub Repository: <https://github.com/kiropisaac/Breast-Cancer-Prediction-Models>

Video Demonstration: <https://youtu.be/JgN4lcL6qaE>