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: *Kirapisaack* **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

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: RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the experimental results, performance evaluation, and interpretation of the machine learning models developed for:

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

The objective is to determine how accurately clinical and pathological variables can be used to predict tumor biology and patient survival, and to evaluate the clinical usefulness of the developed predictive models.

4.2 Data Exploration and Preparation Results

Initial exploratory data analysis revealed several important characteristics of the dataset:

- The dataset contained 2,509 patient records with both clinical and outcome variables.
- A mixture of numerical and categorical features required preprocessing before modelling.
- Some clinical variables contained missing values, which were handled through:

    i). Median imputation for numerical data

    ii). Removal or appropriate handling of missing categorical entries

Feature encoding and scaling ensured compatibility with machine learning algorithms and improved model stability.

4.2.1 Class Distribution

The target variables showed class imbalance, particularly in:

- Certain molecular subtypes
- Survival outcome categories

This imbalance is common in medical datasets and was considered during model evaluation and interpretation, especially when analyzing precision, recall, and F1-score rather than relying only on accuracy.

4.3 Model Training Overview

Three supervised learning algorithms were trained and compared:

- Logistic Regression
- Random Forest Classifier
- Extreme Gradient Boosting (XGBoost)

Separate models were developed for each prediction task to allow task-specific optimization and clearer interpretation of results.

The dataset was divided into:

- Training set for learning model parameters
- Testing set for unbiased performance evaluation

This ensured that reported results reflect true predictive capability rather than memorization of training data.

4.4 Molecular Subtype Prediction Results

4.4.1 Performance Summary

Among the evaluated models:

- Logistic Regression provided moderate performance and strong interpretability but struggled with complex nonlinear relationships.
- Random Forest significantly improved classification accuracy and handled feature interactions effectively.
- XGBoost achieved the highest overall predictive performance, demonstrating superior capability in modelling complex clinical patterns.

4.4.2 Interpretation

High predictive accuracy in subtype classification indicates that:

- Routinely collected clinical and biomarker variables contain meaningful signals related to tumor molecular biology.
- Machine learning can act as a practical alternative when genomic testing is unavailable**.**

This finding supports the feasibility of cost-effective subtype prediction in resource-limited healthcare environments.

4.5 Binary Survival Prediction Results

4.5.1 Performance Summary

For predicting whether a patient is living or deceased**:**

- Logistic Regression provided a reliable baseline.
- Random Forest improved sensitivity to high-risk patients.
- XGBoost again produced the best balance of precision, recall, and F1-score.

4.5.2 Clinical Interpretation

Accurate binary survival prediction enables:

- Early identification of high-risk patients
- Closer clinical monitoring
- More aggressive or personalized treatment planning

This demonstrates the clinical decision-support potential of the developed model.

4.6 Multi-Class Vital Status Prediction Results

4.6.1 Performance Summary

Predicting:

- Living
- Died of disease

- Died of other causes

is inherently more complex due to overlapping clinical characteristics.

Results showed:

- Logistic Regression had limited discrimination between death causes.
- Random Forest improved class separation.
- XGBoost achieved the most consistent multi-class performance, though slightly lower than binary prediction accuracy due to task complexity.

4.6.2 Interpretation

The ability to distinguish cancer-specific mortality from other causes is valuable for:

- Clinical prognosis studies
- Treatment effectiveness evaluation
- Public health and survival research

This confirms the usefulness of multi-class predictive modelling in oncology.

4.7 Comparative Model Analysis

Across all three prediction tasks:

- XGBoost consistently outperformed other algorithms.
- Random Forest ranked second, offering strong interpretability and robustness.
- Logistic Regression served as an interpretable baseline but lacked nonlinear modelling power.

4.7.1 Key Insight

Ensemble learning methods are particularly effective for:

- Complex medical datasets
- Nonlinear feature interactions
- Imbalanced clinical outcomes

This aligns with findings from previous machine learning research in healthcare.

4.8 Clinical and Practical Implications

The developed predictive system demonstrates that:

- Machine learning can approximate molecular subtype information using clinical data.
- Survival outcomes can be predicted with meaningful accuracy.
- Integrated prediction models can support evidence-based oncology decision-making.

Potential Real-World Applications

- Decision support in hospitals lacking genomic testing
- Risk stratification tools for oncologists
- Research support for survival outcome studies

However, the system should be used as a support tool rather than a replacement for medical professionals**.**

4.9 Limitations of the Results

Despite promising performance, several limitations remain:

- Missing data may influence predictive accuracy.
- Dataset population characteristics may limit global generalization.
- Machine learning predictions remain probabilistic rather than deterministic.

Future work should include:

- External clinical validation
- Integration with real hospital systems
- Inclusion of imaging or genomic data for improved accuracy

4.10 Summary

This chapter presented the experimental findings and interpretation of the developed machine learning models. Key conclusions include:

- Clinical data can effectively predict breast cancer molecular subtype and survival.

- XGBoost achieved the best performance across all prediction tasks.
- The system shows strong potential as a cost-effective clinical decision-support tool.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 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.

5.2 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.

5.3 Key Findings

The experimental evaluation produced several important findings:

5.3.1 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.

5.3.2 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.

5.3.3 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**.**

5.4 Contributions of the Study

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

5.4.1 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.

5.4.2 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.

5.5 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**.**

5.6 Recommendations

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

5.6.1 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.

5.6.2 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.

5.7 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.

5.8 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.