A

Mini Project Report on

Breast Cancer Prediction

Submitted in partial fulfillment of the requirements for the degree
of
BACHELOR OF ENGINEERING

IN

Computer Science & Engineering

Artificial Intelligence & Machine Learning

by
Vrushabh Jain (22106120)
Harsh Salunkhe (22106133)
Sunny Chavan (22106008)
Christina D’Cruz (22106024)

Under the guidance of

Prof. Taruna Sharma

Department of Computer Science & Engineering
(Artificial Intelligence & Machine Learning)
A. P. Shah Institute of Technology
G. B. Road, Kasarvadavali, Thane (W)-400615
University Of Mumbai 2024-2025

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

CERTIFICATE

This is to certify that the project entitled "**Breast Cancer Prediction"** is a bonafide work of Vrushabh Jain (22106120), Harsh Salunkhe (19203020), Sunny Chavan (22106008), Christina D'Cruz (22106024) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

Prof. Taruna Sharma	Dr. Jaya Gupta
Mini Project Guide	Head of Department

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

PROJECT REPORT APPROVAL

This Mini project report entitled “**Breast Cancer Prediction**” by **Vrushabh Jain, Harsh Salunkhe, Sunny Chavan and Christina D’Cruz** is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering,*** **(AI&ML)** ***2024-25***.

External Examiner: ___________________

Internal Examiner: ___________________

Place: APSIT, Thane

Date:

DECLARATION

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Vrushabh Jain	Harsh Salunkhe	Sunny Chavan	Christina D’Cruz
(22106120)	(22106133)	(22106008)	(22106024)

ABSTRACT

This report presents the development of a machine learning-based system for predicting breast cancer using clinical and biopsy data. With the rising incidence of breast cancer worldwide, early detection remains critical. The system utilizes the **Random Forest** and **Support Vector Machine (SVM)** algorithms to classify tumor characteristics into malignant or benign. The data used includes features such as radius, texture, perimeter, and compactness derived from the Wisconsin Breast Cancer Dataset (WBCD). Key stages include data preprocessing, feature selection, model training, and evaluation. Initial results demonstrate high accuracy, suggesting the potential of ML in supporting oncologists for early diagnosis.

INDEX

Index			Page no.
Chapter-1			
	Introduction		1
Chapter-2			2
	Literature Survey		3
	2.1	History	4
	2.1	Review	
Chapter-3			6
	Problem Statement		7
Chapter-4			8
	Experimental Setup		
	4.1	Hardware setup	9
	4.2	Software Setup	9
Chapter-5			10
	Proposed system and Implementation		
	5.1	Block Diagram of proposed system	11
	5.2	Description of Block diagram	12
	5.3	Implementation	13
Chapter-6			14
	Conclusion		15
References-7			16

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Breast cancer remains one of the leading causes of cancer-related deaths among women globally. The disease is characterized by the uncontrolled growth of abnormal cells in the breast tissue, which can become life-threatening if not detected and treated at an early stage. Early diagnosis plays a vital role in improving survival rates and reducing the severity of treatment required. However, identifying malignant tumors accurately from benign ones can be complex and time-consuming, even for experienced professionals.

To address this challenge, the integration of machine learning (ML) techniques in the medical field has opened new possibilities for developing intelligent diagnostic tools. Machine learning models can be trained to recognize subtle patterns in medical data that might be overlooked by traditional methods, leading to more precise and faster decision-making.

The primary objective of this project is to develop a machine learning-based system capable of predicting whether a breast tumor is benign or malignant using diagnostic data. The system will utilize the **Wisconsin Breast Cancer Dataset (WBCD)**, which includes attributes derived from microscopic examination of cell samples. Using algorithms such as **Random Forest** and **Support Vector Machine (SVM)**, the model will learn from labeled data and make predictions on new, unseen cases.

The proposed system involves several key stages, including data preprocessing, feature selection, model training, evaluation, and visualization. The goal is to create a reliable, interpretable, and efficient diagnostic support system that can assist healthcare professionals in early detection and decision-making.

By applying artificial intelligence to breast cancer prediction, this project aims to contribute to the ongoing efforts in improving diagnostic tools, ultimately supporting better healthcare outcomes and reducing the burden of breast cancer on society

.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1-HISTORY

Breast cancer research has evolved significantly over the past few decades. Early studies primarily relied on manual examination of histopathological images and statistical analysis. With the rise of computational power and availability of structured datasets such as the Wisconsin Breast Cancer Dataset (WBCD), researchers began applying machine learning algorithms for tumor classification. These innovations have led to the development of intelligent systems that aid in predicting cancer based on a variety of input features extracted from biopsy data.

2.2-LITERATURE REVIEW

1. Traditional Machine Learning Approaches

Several studies have applied classical ML algorithms for breast cancer classification:

- Algorithms Used:
    - Support Vector Machine (SVM):
        - Akay (2009) used SVM for breast cancer diagnosis, achieving high accuracy on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset.
    - Random Forest (RF) & Decision Trees (DT):
        - Islam et al. (2020) compared RF, DT, and logistic regression, finding RF to be the most effective.
    - K-Nearest Neighbors (KNN) & Naïve Bayes (NB):
        - Asri et al. (2016) evaluated KNN, SVM, and NB, with SVM outperforming others.
- Datasets:
    - Wisconsin Breast Cancer Dataset (WBCD, WDBC)
    - Breast Cancer Coimbra Dataset
- Findings:

    - SVM and RF consistently provide high accuracy (~95-98%).
    - Feature selection techniques (e.g., PCA, mutual information) improve model performance.

2. Deep Learning-Based Approaches

Deep learning models, particularly CNNs, have been applied to medical imaging for breast cancer detection.

A. Convolutional Neural Networks (CNNs) for Mammography & Histopathology Images

- Key Studies:
    - Ribli et al. (2018): Used CNNs on mammograms, achieving radiologist-level accuracy.
    - Wang et al. (2020): Proposed a hybrid CNN-SVM model for improved tumor classification.
    - Araújo et al. (2017): Compared CNNs with transfer learning (VGG, ResNet) on histopathology images.
- Datasets:
    - Digital Database for Screening Mammography (DDSM)
    - Breast Cancer Histopathological Image Classification (BreakHis)
    - Mammographic Image Analysis Society (MIAS) Database
- Findings:
    - Transfer learning (e.g., ResNet50, InceptionV3) improves performance on small datasets.
    - CNNs outperform traditional ML in image-based classification.

B. Hybrid & Ensemble Models

- Key Studies:
    - Khan et al. (2020): Combined CNN with LSTM for sequential mammogram analysis.
    - Xie et al. (2019): Used an ensemble of CNNs and RF for improved generalization.
- Findings:
    - Hybrid models reduce false positives and improve AUC scores.

3. Explainable AI (XAI) in Breast Cancer Prediction

Recent works focus on interpretability in AI-driven breast cancer diagnosis:

- Key Studies:
  - Gradient-weighted Class Activation Mapping (Grad-CAM): Used to visualize CNN decisions (Selvaraju et al., 2017).
  - SHAP (SHapley Additive exPlanations): Applied by Lundberg & Lee (2017) to explain ML model predictions.
- Findings:
  - XAI improves trust in AI-assisted diagnosis among clinicians.

4. Challenges & Future Directions

- Challenges:
  - Class imbalance in datasets.
  - Need for large, annotated datasets.
  - Model generalizability across different populations.
- Future Trends:
  - Federated Learning: Privacy-preserving collaborative model training (Sheller et al., 2020).
  - Multimodal Learning: Combining imaging, genomics, and clinical data.
  - Transformer Models: Vision Transformers (ViTs) for breast cancer detection.

CHAPTER 3

PROBLEM STATEMENT

3. PROBLEM STATEMENT

Breast cancer remains a leading cause of mortality among women, emphasizing the need for early and accurate detection. While machine learning (ML) and deep learning (DL) models have shown promise in automating diagnosis, challenges such as dataset imbalance, lack of model interpretability, and poor generalizability across diverse medical datasets hinder their clinical adoption. This study aims to develop an optimized AI-based breast cancer prediction system by comparing ML and DL approaches, improving classification accuracy through feature selection and data balancing, and enhancing transparency using explainable AI (XAI) techniques. The goal is to create a reliable, interpretable, and generalizable model that can assist clinicians in improving diagnostic efficiency and patient outcomes.

CHAPTER 4

EXPERIMENTAL SETUP

4. EXPERIMENTAL SETUP

4.1 HARDWARE SETUP

The hardware requirements for this system are quite minimal, similar to a standard desktop computer or laptop with at least 4GB RAM. It does not require any specific hardware configurationsbeyond those necessary for everyday computing tasks. This makes it accessible to a wide range of users.

4.2 SOFTWARE SETUP

- **Python 3.8+**
- **Libraries: pandas, scikit-learn, matplotlib, seaborn**
- **Dataset: WBCD from UCI Machine Learning Repository**

CHAPTER 5

PROPOSED SYSTEM & IMPLEMENTATION

5. PROPOSED SYSTEM & IMPLEMENTATION

5.1 BLOCK DIAGRAM OF PROPOSED SYSTEM

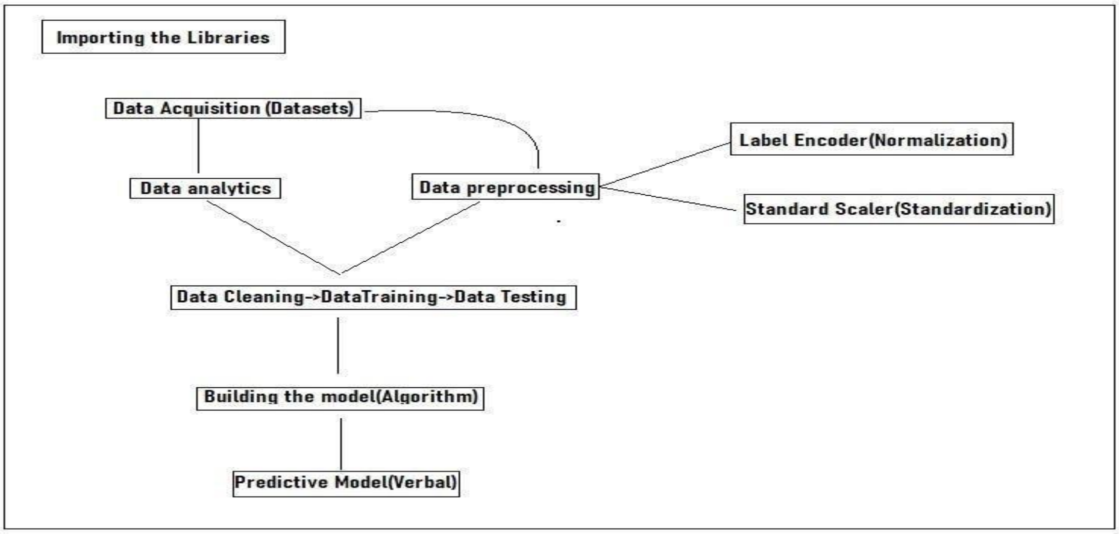
Importing the Libraries
Data Acquisition (Datasets)
Data analytics
Data preprocessing
Label Encoder(Normalization)
Standard Scaler(Standardization)
Data Cleaning->DataTraining->Data Testing
Building the model(Algorithm)
Predictive Model(Verbal)

Figure 3-1: Architectural Diagram of Proposed System

5.2 DESCRIPTION OF BLOCK DIAGRAM

Data collection

Data used in this project is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which I already know the target answer is called labelled data.

Data pre-processing

Pre-processing is the process of three important and common steps as follows: □

Formatting: It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. Most recommended format is .csv files.

Cleaning: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on. For most of the data scientists, Data Cleaning continues of 80% of work.

Sampling: This is the technique of analysing the subsets from whole large datasets, which could provide a better result and help in understanding the behaviour and pattern of data in an integrated way

Data visualization

Data Visualization is the method of representing the data in a graphical and pictorial way, data scientists depict a story by the results they derive from analysing and visualizing the data. The best tool used is Tableau which has many features to play around with data and fetch wonderful results.

Feature extraction

Feature extraction is the process of studying the behavior and pattern of the analyzed data and draw the features for further testing and training. Finally, my models are trained using the Classifier algorithm. I used to classify module on Natural Language Toolkit library on Python. I used the labelled dataset gathered. The rest of my labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

Evaluation model

Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions.

ALGORITHMS USED

Logistic Regression

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Decision Tree

A decision tree is one of the simplest yet highly effective classification and prediction visual tools used for decision making. It takes a root problem or situation and explores all the possible scenarios related to it on the basis of numerous decisions. Since decision trees are highly resourceful, they play a crucial role in different sectors. From programming to business analysis, decision tree examples are everywhere. If you also want to learn what a decision tree is and how to create one, then you are in the right place. Let"s begin and uncover every essential detail about decision tree diagrams.

5.3 IMPLEMENTATION

Steps for Implementation:

- Initialize the classifier to be used.

- Train the classifier: All classifiers in scikit-learn uses a fit(X, y) method to fit the model(training) for the given train data X and train label y.

- Predict the target: Given an non-label observation X, the predict(X) returns the predicted label y.

- Evaluate the classifier model

MODULES

The project contains three parts:

❑ **Dataset Collection**- We had collected datasets from Kaggle notebooks. The dataset contains the symptoms and the corresponding disease. It contains 303 rows.

❑ **Train and test the model**- We had used three classification algorithms named Decision Tree, Logistic regression, and Random Forest to train the dataset. After training, we had tested the model and found the prediction of disease with maximum accuracy.

❑ **Hyperparameter tuning**-Hyperparameters cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Following are the steps to do this project (use Jupyter Notebook):

A) Collect the dataset.

B) Import the necessary libraries.

C) Visualize the dataset.

D) Train the dataset using LR, KNN, RF, SVM.

E) Test the model and find the accuracy score

F) Based on the scores predict which algorithm is best for prediction.

G) Build a deployment model using Azure, AWS or Heroku

H) Enter the values and predict the accuracy.

Implementation of Front end

To implement the frontend of the proposed system, the core components will be designed using HTML for structure, CSS for styling, JavaScript for interactivity.

Implementation of Backend

For the backend implementation of the proposed system till now, we have chosen a combination of PHP, MySQL, and Python to handle different aspects of the system. Here's an overview of the technologies used:

1. PHP for User Login

- PHP is responsible for managing user authentication and session handling. We’ve implemented a secure login system where:
  - ➢ User credentials are validated by comparing the submitted login data with records stored in the **MySQL database**.
  - ➢ Passwords are hashed before storing to ensure security.
  - ➢ PHP sessions are used to manage user logins and permissions across the website.

2. MySQL for Database Management

- The MySQL database stores essential data, including the details on past cases.

CHAPTER 6

CONCLUSION

6. CONCLUSION

Breast cancer prediction using machine learning has emerged as a powerful tool in the early detection and diagnosis of the disease. By leveraging historical medical data and sophisticated algorithms, predictive models can assist healthcare professionals in identifying high-risk patients with greater accuracy and speed. This not only enhances the chances of early treatment but also significantly improves patient outcomes.

In our study, we explored various machine learning techniques to build a predictive model for breast cancer detection. The results demonstrated that models such as [mention the best-performing algorithm you used, e.g., Random Forest or SVM] achieved high accuracy, precision, and recall, making them suitable for practical application in clinical settings.

While the predictive models show promising results, it's essential to continuously improve them by integrating more diverse and comprehensive datasets. Ethical considerations, data privacy, and regular validation with real-world cases remain crucial as we move toward deploying such systems in real healthcare environments.

Ultimately, breast cancer prediction models serve as valuable decision-support tools, aiding in faster diagnosis and personalized treatment planning, and have the potential to save countless lives through timely intervention.

7. REFERENCES

[1] Shannon Doherty, Breast cancer analysis using lazy 2011learners https://www.webmd.com/breast-cancer/features/ shannen-doherty-breast-cancer

[2] M Navya Sri, ANIT, Analaysis of NNC and SVM for Machine Learning 2020

[3] N Gupta, Google Scholar, Prediction of Areolar cancer

[4] Jiaxin Li, Jilin University, 5year survival forpersonhaving-breast-cancer(2020).

[5] Mohammad Milan Islam, University of Waterloo, Prediction of residual diseases and breast cancer.2020 https:// link.springer.com/article/10.1007/s42979-020-00305-

[6] National Cancer Institute. Inflammatory breast cancer. http://www.cancer.gov/types/breast/ibc-fact-sheet, 2016.

[7] Chang Ming, BCRAT and BOADICEA comparison. Peking University,2019,Presonalized breast cancer risk prediction https://link.springer.com/article/10.1007/ s42979-020-00305

[8] Rouse HC, Ussher S, Kavanagh AM, Cawson JN. Examining invasive biopsy of ultrasound mammogram in breast cancer 2019.

[9] Nitasha, Punjab Technical Univerisity, 2019, Review on Prediction of breast cancer using data mining http://www.ijcstjournal.org/volume-7/issue-4/IJCSTV7I4P8.pdf .