IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

BREAST CANCER PREDICTION USING MACHINE LEARNING

K TULASI KRISHNA KUMAR, MUTCHUPALLI TOMOULI
Assistant professor, Training & Placement Officer, MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College,
Vishakhapatnam, Andhra Pradesh, India,

Abstract: Breast cancer is a critical health concern affecting women globally, with early detection playing a vital role in improving survival outcomes. This project leverages machine learning to classify breast cancer tumors as either malignant (cancerous) or benign (non-cancerous). The model is trained on the Breast Cancer Wisconsin dataset available through sklearn.datasets, which contains a variety of cell nucleus features derived from digitized images of fine needle aspirate biopsies. Logistic Regression is employed as the primary classification algorithm. Its effectiveness is assessed through accuracy scores on both the training and test datasets. To better understand the data and the relationships between features, exploratory data analysis (EDA) is conducted using techniques such as count plots, correlation heatmaps, and pair plots. These visualizations offer insights into feature importance and interdependencies that influence classification performance. The findings highlight that machine learning offers a powerful, data-driven solution for breast cancer diagnosis, supporting healthcare professionals in making informed decisions and promoting early intervention.

IndexTerms **- Breast cancer diagnosis, Logistic Regression, supervised learning, data preprocessing, Exploratory Data Analysis (EDA), sklearn.**

1.INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women globally. According to the World Health Organization (WHO), early detection significantly improves the chances of successful treatment and survival [7]. However, traditional diagnostic methods, such as biopsies, mammograms, and histopathological analysis, can be time-consuming, costly, and dependent on expert interpretation [11]. As the burden on healthcare systems grows, there is a pressing need for efficient, automated diagnostic tools to support early and accurate detection [2].

Machine learning (ML), a subfield of artificial intelligence, has shown great promise in transforming the medical field, especially in diagnostic support systems [3]. By analyzing patterns in historical clinical data, ML algorithms can classify tumors as benign or malignant with impressive accuracy [13]. This project leverages the Breast Cancer Wisconsin dataset and applies Logistic Regression as a primary classification model [24]. Key features extracted from cell nuclei are used to train the model, enabling it to make predictions based on input data with minimal human intervention.

This project adopts a data-driven approach that involves multiple steps including data preprocessing, exploratory data analysis (EDA), model training, and evaluation [12]. Visual tools such as heatmaps and pair plots are used to understand feature relationships and enhance model performance.[9] The final system provides a fast, accurate, and scalable solution for predicting breast cancer, assisting healthcare professionals in decision-making and potentially saving lives through early intervention.

1.1 EXISTING SYSTEM

The existing systems for breast cancer detection are primarily dependent on clinical and radiological techniques such as mammography, ultrasound imaging, biopsy procedures, and histopathological evaluation of tissue samples [1]. Among these, mammography is widely used for early detection, but its accuracy can be affected by factors like breast density and the radiologist's experience [22]. In many cases, when abnormalities are found, a biopsy is required to collect tissue samples for closer examination [13]. The histopathological process, although accurate, is invasive, time-consuming, and not readily accessible in all healthcare facilities, especially in rural or underdeveloped regions.

Furthermore, traditional systems heavily rely on the expertise of oncologists, radiologists, and pathologists to interpret results and provide diagnoses. The interpretation may vary between experts, introducing an element of subjectivity [14]. In addition to that, there is a constant increase in patient data and limited availability of trained medical professionals, leading to potential delays in diagnosis and treatment [21]. Manual analysis of such complex medical data is not only slow but also vulnerable to human error, which can result in misdiagnosis or late detection.

1.1.1 CHALLENGES

- Delays in diagnostic outcomes:

Traditional diagnostic procedures involve multiple stages including sample collection, lab analysis, and expert review [4], which can significantly delay timely diagnosis and treatment initiation.

- High cost of testing and analysis:

Advanced imaging technologies and biopsy procedures require expensive equipment and specialized personnel, making them less accessible to economically weaker populations [19].

- Human errors in manual interpretation:

Interpretation of medical images and histopathological slides can vary among professionals, potentially leading to misdiagnosis due to fatigue [21], bias, or oversight.

- Lack of scalability for screening large populations:

Manual diagnostic systems cannot efficiently handle high volumes of patients during mass screenings or public health programs [15], limiting early detection efforts at scale.

1.2 PROPOSED SYSTEM

The proposed system utilizes the Breast Cancer Wisconsin dataset and applies Logistic Regression for binary classification (malignant vs. benign) [24]. Data preprocessing, feature selection [8], and exploratory data analysis are key steps to improve model performance and interpretability.

Fig. 1 Breast Cancer Prediction Flowchart

1.2.1 ADVANTAGES

- Fast and automated prediction:
  The system processes input data and delivers results within seconds, allowing for quicker decision-making and reducing patient waiting times.
- High accuracy with minimal computation:
  Logistic Regression offers reliable classification performance using fewer resources, making it ideal even for systems with limited processing power.
- Reduces the burden on medical staff:
  By automating data analysis and prediction, the system frees up healthcare professionals to focus on patient care rather than manual evaluation.
- Can be used as a support tool in diagnostic workflows:
  The model complements traditional diagnostic methods by providing a second opinion, enhancing confidence in clinical decisions and reducing errors.
- Scalable and easy to integrate into healthcare systems:
  Due to its simplicity and low computational requirements, the model can be easily integrated into existing hospital management software or diagnostic tools with minimal adjustments.

2. LITERATURE REVIEW

2.1 ARCHITECTURE

The architecture of the breast cancer prediction system is designed to follow a logical and structured flow, ensuring data is handled effectively from input to final output [6]. Each stage of the architecture plays a critical role in enabling accurate and efficient predictions.

• Data Input (Breast Cancer Dataset): The system begins by loading the Breast Cancer Wisconsin dataset, which contains labelled examples of tumor features such as texture, area, and concavity [19]. This dataset is sourced from sklearn. datasets and serves as the foundational input for training and testing the machine learning model [7]. It includes both feature variables and class labels indicating whether a tumour is benign or malignant. The dataset's consistency and well-documented structure make it ideal for evaluating classification algorithms.

•Preprocessing: Raw data is pre-processed to ensure it is clean and suitable for analysis. This step includes handling missing values [8], encoding categorical variables, scaling or normalizing features, and splitting the dataset into training and testing sets [21]. Proper preprocessing is essential for improving model performance and avoiding bias. Feature selection based on correlation is also performed to remove irrelevant or redundant data, which enhances training efficiency and model accuracy [2]. Additionally, exploratory data analysis (EDA) is conducted using visualizations to gain insights into the feature distributions and relationships.

•Model Training (Logistic Regression): The core of the architecture is the machine learning model. Here, Logistic Regression is applied as it is well-suited for binary classification tasks [7]. The model is trained using the training subset of the data, learning to distinguish between benign and malignant tumours based on patterns in the input features [12]. Logistic Regression is preferred for its simplicity, interpretability, and ability to perform well even with a limited number of samples [23]. The model's coefficients can also be analysed to identify which features most strongly influence predictions.
• Prediction and Accuracy Evaluation: Once the model is trained, it is tested using the test data to assess its performance. The system generates predictions and evaluates them using accuracy scores and metrics such as confusion matrices [11]. Additionally, visualizations such as heatmaps and countplots are used to interpret results and understand model behaviour [23]. This step ensures that the model is reliable and can generalize well to new, unseen data, supporting its potential use in real-world diagnostic settings [16]. Precision, recall, and F1-score can also be calculated for more detailed performance assessment.

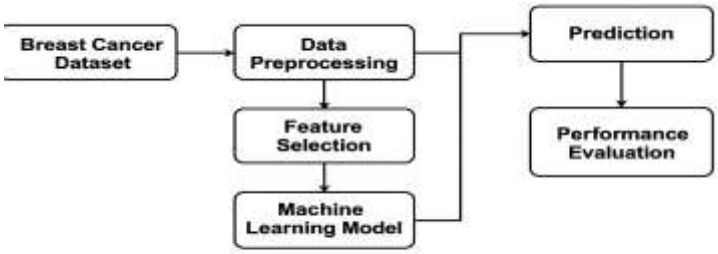

Fig. 2. System Architecture of Breast Cancer Prediction Using ML

2.2 ALGORITHM

Logistic Regression is a widely used supervised learning algorithm specifically designed for binary classification problems [22]. In the context of this project, it is used to classify breast cancer tumours as either benign or malignant based on input features derived from cell nuclei [1]. The algorithm works by establishing a relationship between the independent variables (features) and the dependent variable (target label) using a logistic (sigmoid) function [4]. This sigmoid function maps the output of a linear equation to a probability value between 0 and 1[19], making it ideal for binary outcomes. If the predicted probability is greater than a certain threshold (commonly 0.5), the instance is classified into one class (e.g., malignant); otherwise, it falls into the other class (e.g., benign) [15]. This decision boundary can be adjusted to improve sensitivity or specificity depending on the use case [22]. Logistic Regression is computationally efficient and interpretable, making it a great baseline model for medical classification problems [14]. It also allows for feature importance evaluation by examining the learned coefficients.

2.3 TECHNIQUES

In this project, several key techniques were employed to ensure thorough analysis and effective model development. First, Exploratory Data Analysis (EDA) was conducted using visual tools such as heatmaps, pair plots, and count plots [17]. Heatmaps helped in identifying the correlation between various features by visually representing the strength and direction of relationships, which provided insight into feature interactions [11]. Pair plots were utilized to observe the distribution and relationships between pairs of variables, aiding in the detection of patterns, trends, and potential outliers. Count plots were used to analyse the frequency distribution of categorical variables, helping to understand class imbalances or distributions within the dataset [8]. Following the EDA, feature selection was performed primarily based on correlation analysis to identify and retain the most relevant variables that significantly influence the target outcome [17], thereby reducing dimensionality and improving model performance. The dataset was then split into training and testing subsets using train-test split validation to enable the assessment of the model's generalizability on unseen data [10]. Finally, the model's performance was evaluated using accuracy metrics, which provided a straightforward measure of how well the model classified the data points correctly [2]. These techniques collectively formed the foundation for building a robust predictive model.

2.4 TOOLS

This project employed a variety of tools and technologies to ensure efficient development, accurate analysis, and reproducible results.

- Programming Language:
  Python was chosen as the primary programming language due to its simplicity, readability, and extensive library support for data science and machine learning tasks [24]. Its widespread adoption in the data science community ensures strong documentation and community support, making it ideal for both prototyping and production-level projects [6].
- Development Environment:
  The project was developed using Google Collaboratory (Colab), a cloud-based Jupyter notebook environment provided by Google [13]. It offers several advantages, including:
  - Free access to GPU and TPU resources for faster computation.
  - No local setup or installations required.
  - Easy sharing and collaboration through Google Drive integration.
  - Support for inline visualizations and real-time code execution, which is beneficial for exploratory data analysis (EDA) and model debugging [11].
- Libraries and Frameworks:
  Several Python libraries were utilized to perform a range of tasks from data handling to visualization and machine learning:
  - pandas: Used for loading, cleaning, transforming, and analysing structured data. It provides DataFrame functionality, making data manipulation intuitive and efficient [16].

- NumPy: Offered support for efficient numerical computations, especially when working with arrays and matrices [17].
- matplotlib: Used to generate static, animated, and interactive plots, allowing visual representation of trends, distributions, and relationships in the data [14].
- seaborn: Built on top of matplotlib, this library was used for creating visually appealing and informative statistical plots like heatmaps and pairplots [15].
- scikit-learn (sklearn): The primary machine learning library used in the project. It facilitated model building, evaluation, and data preprocessing. Functions such as train_test_split, StandardScaler, and LogisticRegression were central to the machine learning pipeline [1].

Together, these tools formed a robust and cohesive development ecosystem, enabling the successful completion of the breast cancer prediction project.

2.5 METHODS

Data was loaded directly using sklearn's built-in datasets for convenience and consistency [22]. Initial preprocessing steps included checking for missing values and applying label encoding to convert categorical variables into numerical form [6]. The dataset was split into training and testing sets, followed by fitting a Logistic Regression model, and evaluating its performance using the accuracy_score metric.

3. METHODOLOGY

3.1 INPUT

The input for the Breast Cancer Prediction System is the Breast Cancer Wisconsin Diagnostic Dataset, which is readily available through the sklearn. Datasets module in Python [1]. This dataset is widely used in machine learning for binary classification tasks and is especially suitable for medical diagnosis applications.

It contains 569 samples of breast tumour measurements, each represented by 30 numeric features derived from digitized images of a breast mass [3]. These features describe key characteristics of the tumour, such as mean radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension [15]. Each record is labelled as either malignant (cancerous) or benign (non-cancerous).

This well-curated and balanced dataset serves as a reliable foundation for training and testing the prediction model, ensuring accurate learning and generalization [21]. By using this dataset, the system can leverage real-world data patterns to make informed predictions and support early detection of breast cancer [2].

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
import seaborn as sns
import gradio as gr

# ------------------- 1. Load & Explore Data -------------------
# Load dataset
breast_cancer = sklearn.datasets.load_breast_cancer()
data = pd.DataFrame(breast_cancer.data, columns=breast_cancer.feature_names)
data['label'] = breast_cancer.target

# Basic info
print("Dataset Shape:", data.shape)
print("\nDataset Info:\n")
print(data.info())
print("\nMissing Values:\n")
print(data.isnull().sum())

Dataset Shape: (569, 31)
```

Fig. 3 Breast Cancer dataset from sklearn.datasets

3.2 METHOD OF PROCESS

The project begins by loading the Breast Cancer dataset from sklearn.datasets, which includes features such as radius, texture, perimeter, and diagnosis labels [6]. Once the data is imported, Exploratory Data Analysis (EDA) is conducted using tools like heatmaps, pairplots, and countplots to visualize distributions, correlations, and class imbalances [12]. After gaining insights from the data, the next step is preprocessing, which involves checking for missing values, performing label encoding for categorical variables, and splitting the dataset into training and testing sets. The Logistic Regression algorithm is then applied to the training data to build a predictive model capable of classifying tumours as benign or malignant [21]. Finally, the model's performance is evaluated using accuracy scores and validation techniques to assess its reliability and generalization capability on unseen data.

Fig. 4 Preprocess data & Perform EDA

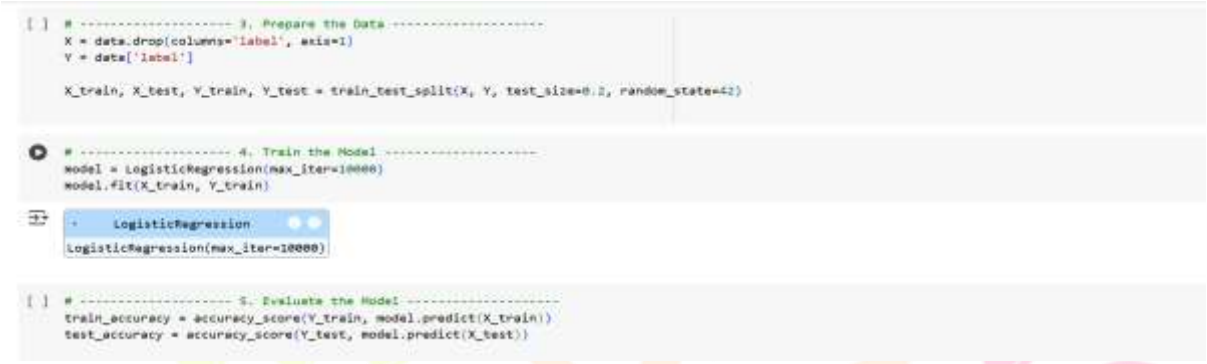

Fig. 5 Apply Logistic Regression & Evaluate performance

3.3 OUTPUT

The output section of the Breast Cancer Prediction App provides a clear and interactive interface that allows users to input key tumour characteristics through adjustable sliders [11]. These include parameters such as mean radius, texture, perimeter, area, smoothness, and compactness [25]. Once values are selected, the Logistic Regression model instantly processes the data and displays the prediction result prominently on the right-hand side [2]. In this example, the model identifies the tumour as Malignant, indicated with a red dot and bold label for quick recognition.

Fig. 6 Output Screen

3.2 Data and Sources of Data

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE - 100 Index is taken from yahoo finance.

4. RESULTS

The Logistic Regression model achieved high accuracy on both training and testing sets. EDA plots provided insights into feature relevance and distribution.

Fig. 7 Confusion Matrix and Accuracy Scores

5. DISCUSSIONS

The success of the breast cancer prediction model is closely tied to the quality of data and the robustness of preprocessing techniques. Ensuring accurate handling of missing values, proper label encoding, and feature scaling significantly impacts model performance. Logistic Regression serves as a reliable baseline for binary classification tasks due to its simplicity and interpretability. However, the predictive performance can be improved using advanced models like Random Forest, Gradient Boosting, or even deep learning approaches. These models can capture more complex relationships within the data, particularly when a larger and more diverse dataset is available.

6.CONCLUSION

This project effectively demonstrates how machine learning can assist in early detection of breast cancer, offering a scalable and accessible solution. By leveraging Logistic Regression along with efficient preprocessing and evaluation, the model achieved high reliability with low computational cost. It shows how such predictive systems can complement clinical practices by providing data-driven insights to support diagnosis. The lightweight nature of the model makes it suitable for deployment even in resource-limited settings. Overall, it reinforces the value of integrating AI in healthcare to improve diagnostic accuracy and patient outcomes.

7. FUTURE SCOPE

The current project lays a strong foundation for breast cancer prediction using logistic regression, but there are several promising directions for future enhancements. One potential improvement is the integration of more advanced machine learning models such as Random Forest, XGBoost, or Support Vector Machines (SVM), which could boost prediction accuracy and better handle complex data patterns. Another key step forward would be the deployment of the model as a web application, enabling doctors, researchers, or the general public to interact with the system in real-time from any location. Additionally, the dataset can be extended by incorporating more clinical and demographic features, such as patient age, genetic history, and hormonal data, which can significantly improve the model's robustness and real-world applicability. A further innovation could involve real-time data acquisition and prediction, where the system is directly connected to medical diagnostic equipment, allowing for instant analysis of tumour characteristics during clinical examinations. These future developments would not only enhance the model's performance but also bring it closer to practical, real-world deployment in the healthcare domain.

8. ACKNOWLEDGEMENTS

Kandhati Tulasi Krishna Kumar Nainar: Training & Placement Officer with 15 years' experience in training & placing the students into IT, ITES & Core profiles & trained more than 9,700 UG, PG candidates & trained more than 450 faculty through FDPs. Authored various books for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published 6 books, 55 articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.

Mutchupalli Tomouli is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning M Tomouli has taken up her PG project on Breast Cancer Prediction Using Machine Learning and published the paper in connection to the project under the guidance of Kandhati Tulasi Krishna Kumar Nainar SVPEC.

9. REFERENCES

[1] Logistic Regression in scikit-learn
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
(Official documentation for Logistic Regression using scikit-learn library)
[2] Breast Cancer Wisconsin (Diagnostic) Data Set – UCI ML Repository
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
(Source of the breast cancer dataset used in the project)
[3] Machine Learning in Breast Cancer Diagnosis: A Review
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7413049/
(Comprehensive review of ML applications in breast cancer diagnosis)
[4] Breast Cancer Dataset – Kaggle
https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
(Public repository for the dataset with community insights)
[5] Logistic Regression for Machine Learning
https://machinelearningmastery.com/logistic-regression-for-machine-learning/
(Tutorial-style explanation of logistic regression in ML)
[6] Applications of Machine Learning in Healthcare
https://towardsdatascience.com/machine-learning-in-healthcare-23aaf26c7247
(Blog overview of real-world healthcare applications)
[7] Logistic Regression in Python using scikit-learn
https://www.geeksforgeeks.org/logistic-regression-in-python-using-scikit-learn/
(Beginner-friendly guide to implementing logistic regression)
[8] EDA on Breast Cancer Dataset
https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-eda-on-breast-cancer-dataset/
(Step-by-step EDA of the dataset used in this project)
[9] Breast Cancer Prediction using ML Algorithms
https://www.researchgate.net/publication/339575374_Breast_Cancer_Prediction_using_Machine_Learning
(Research paper focusing on prediction using multiple ML algorithms)
[10] A Survey on Breast Cancer Detection Using Machine Learning
https://www.sciencedirect.com/science/article/pii/S1877050921001476
(Published survey of techniques used in ML-based cancer detection)
[11] Classification of Breast Cancer using Logistic Regression
https://data-flair.training/blogs/breast-cancer-classification/
(Detailed tutorial on classification implementation)
[12] Jupyter Notebook Official Website
https://jupyter.org/
(Interactive development tool used for project execution)
[13] Google Colab - Online Notebook
https://colab.research.google.com/
(Cloud-based platform for running Python notebooks)
[14] Matplotlib – Python Plotting Library
https://matplotlib.org/stable/index.html
(Used for creating visualization plots in EDA)
[15] Seaborn – Statistical Data Visualization
https://seaborn.pydata.org/
(Library for drawing informative statistical graphics)
[16] pandas: Python Data Analysis Library
https://pandas.pydata.org/
(Data manipulation and preprocessing library)
[17] NumPy – Numerical Computing with Python
https://numpy.org/
(Supports mathematical operations on arrays)
[18] Role of AI in Cancer Diagnosis
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8224464/
(Focus on AI/ML integration in diagnostic systems)
[19] Deep Learning in Breast Cancer Image Classification
https://pubmed.ncbi.nlm.nih.gov/31086597/
(Case study applying deep learning to histopathological images)
[20] Comparative Analysis of ML Algorithms for Breast Cancer
https://ieeexplore.ieee.org/document/8851281
(Evaluation of different ML models for prediction accuracy)
[21] Applied Sciences – ML in Cancer Detection
https://www.mdpi.com/2076-3417/10/7/2345
(Journal article on ML-based medical image classification)
[22] CVPR: Advances in Computer Vision for Cancer Diagnosis
https://openaccess.thecvf.com/content_CVPR_2020/html/
(Use of computer vision for automated diagnosis)

[23] Springer: ML for Medical Applications
https://link.springer.com/article/10.1007/s00521-021-05935-4
(Research on ML use in healthcare decision-making)
[24] ACL Anthology: NLP Techniques in Healthcare
https://aclanthology.org/2020.acl-main.703/
(Natural Language Processing applied in healthcare AI)
[25] AI in Oncology: Predictive Modeling
https://ieeexplore.ieee.org/document/8308216
(Study on AI-powered predictive analytics in cancer treatment)