**Project Report on**

**Breast Cancer Prediction Using Machine Learning**

**A Supervised Learning Approach using Random Forest**

Submitted by:

**Nishant Tripathi**

**B.Tech – Computer Science & Engineering**

**KIET GROUP OF INSTITUTION**

**Batch: 2024– 2028**

**Divyansh Tripathi**

**B.Tech – Chemical Engineering**

**Indian Institute of Technology, Kanpur**

**Batch: 2022 – 2026**

This machine learning project was collaboratively developed by two engineering students from different institutions as part of their personal AI/ML portfolio. The project aims to predict the nature of breast tumors using supervised learning for real-world applications.

**Acknowledgement**

****

I would especially like to thank my brother, Divyansh Tripathi, for his constant support, guidance, and collaboration. From helping me understand complex concepts to debugging errors and brainstorming solutions, his contribution has been invaluable. Working with him made this project not only more educational but also incredibly enjoyable and motivating.

I am also grateful to the open-source community and the developers behind Python, scikit-learn, pandas, and related tools, whose resources made it possible to bring this idea to life.

**Index**

* Introduction
* Problem Statement
* Objectives of the Project
* Dataset Overview
* Tools & Technologies Used
* Data Preprocessing
* Model Selection & Training
* Evaluation Metrics
* Results & Performance Analysis
* User Input & Prediction System
* Conclusion
* References

Introduction:

Breast cancer is one of the most common types of cancer affecting women worldwide. The early detection and diagnosis of breast cancer play a crucial role in improving survival rates and enabling timely medical intervention. However, traditional diagnostic methods often rely on manual analysis, which can be time-consuming and prone to human error.

With the rise of machine learning and data science, there is a growing opportunity to build predictive models that can assist in the detection of breast cancer using historical patient data. By leveraging medical datasets and applying classification algorithms, machine learning models can provide reliable predictions that support doctors and researchers in clinical decision-making.

This project focuses on building a machine learning model that can predict whether a breast tumor is malignant (cancerous) or benign (non-cancerous) based on 30 numerical features extracted from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Problem Statement:

Breast cancer, if not detected early, can lead to severe health complications and even death. Traditional diagnostic processes involve biopsies and manual examination, which, although effective, can be slow and may sometimes lead to misdiagnosis due to human error or limited medical resources.

Given the availability of large medical datasets and advancements in machine learning, it is possible to automate parts of the diagnostic process. Machine learning models can be trained to recognize patterns and relationships between medical features of a tumor and its classification as malignant or benign.

The primary problem this project addresses is:

> How can we build a reliable and accurate machine learning model that can classify a breast tumor as malignant or benign using data from diagnostic reports?

This involves:

- Understanding the structure and content of a real-world medical dataset

- Preprocessing and scaling features appropriately

- Selecting and training the right classification algorithm

- Validating the model using real, unseen test data

- Creating a way for users to input new values and receive live predictions

By solving this problem, the model can serve as a supportive diagnostic tool to help medical professionals or educational demonstrations in ML-based healthcare solutions.

Objectives of the Project:

The main objective of this project is to build a machine learning-based system that can accurately predict whether a breast tumor is malignant or benign, using real-world medical data. The model is designed to assist in early detection of breast cancer and demonstrate the practical application of machine learning in healthcare.

The project also aims to give the developers hands-on experience with real datasets, end-to-end ML pipeline development, and user interaction through prediction systems.

Dataset Overview:

- To understand and explore the Breast Cancer Wisconsin dataset

- To analyze the statistical distribution and structure of all 30 features

- To visualize class distribution (benign vs malignant)

Tools & Technologies Used:

- Programming Language: Python

- Libraries: scikit-learn, pandas, NumPy, matplotlib, seaborn

- Algorithms: Random Forest Classifier, Logistic Regression

- Environment: Google Colab

Data Preprocessing:

- To handle raw data from the dataset

- To scale feature values using StandardScaler

- To prepare the data for model training and testing using `train\_test\_split`

Model Selection & Training:

- To evaluate different ML classification models

- To choose the most accurate and efficient model (Random Forest)

- To train the model on preprocessed training data

Evaluation Metrics:

- To test model accuracy using unseen test data

- To use confusion matrix, precision, recall, and F1-score for evaluation

- To visualize performance using heatmaps

Results & Performance Analysis:

- To interpret classification results with real examples

- To check for misclassifications and understand their causes

- To fine-tune the model for better generalization

User Input & Prediction System:

- To allow users to input 30 feature values (from a medical report)

- To generate real-time predictions of Malignant (0) or Benign (1)

- To show prediction confidence (%) for interpretability

References:

1. W. H. Wolberg, W. N. Street, and O. L. Mangasarian

Machine learning techniques to diagnose breast cancer from fine-needle aspirates,

Cancer Letters, vol. 77, no. 2–3, pp. 163–171, 1994.

2. N. Patil, P. Pawar, and N. Narkhede,

Breast Cancer Prediction Using Machine Learning Techniques,

International Research Journal of Engineering and Technology (IRJET),

vol. 6, no. 5, pp. 3749–3754, May 2019.

3. Scikit-learn: Machine Learning in Python

Pedregosa et al., Journal of Machine Learning Research, 2011

7. Python Libraries: pandas, NumPy, matplotlib, seaborn

https://pandas.pydata.org/

https://numpy.org/

https://matplotlib.org/

https://seaborn.pydata.org/