RESEARCH ARTICLE OPEN ACCESS

Breast Cancer Detection using Machine Learning

Muskan.k(1), Dr.Kumar Siddamallappa. U(2), Anusha Jajur. J(3)

(1) Student, DOS in Computer Application, Davangere University,Shivagangothri, Davangere-577007 karnataka,India

(2) Assistant Professor, DOS in Computer Application, Davangere University,Shivagangothri, Davangere-577007 karnataka,India

(3)Research Scholar, DOS in Computer Science, Davangere University,Shivagangothri, Davangere-577007 karnataka,India

ABSTRACT

Breast cancer is one of the most common and life-threatening diseases affecting women worldwide. Early and accurate diagnosis is critical for effective treatment and increased survival rates. This project focuses on the analysis of a breast cancer dataset using machine learning techniques to predict whether a tumor is malignant or benign based on various physical characteristics. The dataset, sourced from the UCI Machine Learning Repository, includes 569 records with 33 features derived from digitized images of fine needle aspirate (FNA) of breast masses. After eliminating irrelevant or empty columns, such as the 'Unnamed: 32' column, the data was cleaned, normalized, and preprocessed for modeling. Exploratory Data Analysis (EDA) was conducted to visualize feature distributions and detect correlations between variables, providing insights into which features are most indicative of malignancy. Feature selection techniques were used to retain the most relevant predictors, enhancing model performance and interpretability. Several classification algorithms were evaluated, including Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. These models were assessed using performance metrics such as accuracy, precision, recall, and F1 score. The results revealed that models like SVM and Random Forest achieved high accuracy and reliable diagnostic capability, with specific features such as radius_mean, texture_mean, and concavity_worst playing a significant role in classification.Checked accuracy on the test dataset as 94% accuracy. Simulated development by running the saved model with new unseen data to confirm the stability.

Keywords: EDA, KNN, SVM, FNA.

1.INTRODUCTION

Breast cancer is one of the most widespread cancers globally, and early detection is key to saving lives. Traditional diagnosis relies heavily on imaging and clinical expertise, but human judgment can among specialists. Machine learning offers a valuable support system by learning from large datasets of past cases. It delivers fast, consistent predictions that help doctors make more informed decisions. While it doesn't replace medical professionals, it enhances diagnostic accuracy and efficiency, making healthcare more reliable and accessible.The treatment depends on the cancer type and stage may involve surgery We want a simple system that students and beginners can understand and still use in practice. The aim is to make a working demo that runs locally in a browser and predicts using a trained model. This concise overview offers insight into breast cancer's nature, symptoms, risks, and management options that can help with early detection and effective treatment.

1. Breast Cancer Diagnosis: Applying machine learning techniques to predict whether a breast tumor is malignant or benign based on diagnostic imaging features.
2. Causes of Breast Cancer: Genetic mutations, hormonal imbalances, environmental exposures, and lifestyle factors contributing to the development of breast cancer.
3. Symptoms of Breast Cancer: Physical signs such as lumps in the breast, changes in breast shape, nipple discharge, and skin dimpling.
4. Medical Reasons: Abnormal cell growth, mutations in BRCA1/BRCA2 genes, and irregularities in tissue structures detected through imaging and biopsy.
5. Research and Risk: Ongoing studies analyzing the correlation between age, family history, hormonal exposure, and breast cancer risk.
6. Datasets Available: Public datasets like the Wisconsin Breast Cancer dataset from UCI Repository containing tumor measurements for model training and evaluation.
7. Safety Measures and Preventions: Regular screening, mammography, genetic testing, and lifestyle adjustments to lower the risk and ensure early detection.

.

2. *LITERATURE SURVEY*

Many studies have applied machine learning to the Wisconsin dataset. Logistic Regression often performs strongly because the data is clean and linearly separable to some extent.

Support Vector Machines (SVM) also show high accuracy. Random Forests and Gradient Boosting are powerful when we want non-linear decision boundaries.

SL. No	Title	Author and Publisher	Key Points	Remark
1	Machine Learning Approaches for Breast Cancer Classification	Md.Sifat Momen et al. (1) 2023	Applied SVM, KNN , and Random Forest classification. - Emphasized feature importance and dimensionality reduction.	Offers comparative insights on ML model effectiveness in breast cancer diagnosis.
2	Deep Learning Techniques for Breast Cancer Detection	Ayesha Khan et al.(2)2022	- Use of CNN models with mammographic images. - Highlighted benefits of deep learning in image-based diagnosis.	Shows potential of DL models for image classification in medical imaging.
3	Predictive Analytics in Breast Cancer Diagnosis Using Machine Learning	R. S. Rajput, IEEE et al.(3) 2021	- Used data preprocessing, feature selection, and multiple ML classifiers. - Performance metrics such as accuracy and F1-score compared.	Demonstrates structured ML pipeline for breast cancer prediction.
4	A Review of Machine Learning Models for Breast Cancer Detection	Priya R.et al, (4)2020	- Reviewed various algorithms: Naïve Bayes, Decision Tree, Logistic Regression. - Discussed challenges like over fitting and data imbalance.	Helpful for model selection and identifying research gaps.
5	Early Detection of Breast Cancer Using Hybrid ML Models	Meenakshi Sundaram et al(5). 2019	- Combined ML algorithms for ensemble learning. - Focus on improved sensitivity and reduced false negatives.	Promotes hybrid models as effective alternatives to single classifiers.

3.PROPOSED SYSTEM

The proposed system for Breast Cancer Prediction involves the development and deployment of a machine learning-based diagnostic model that utilizes detailed patient data—such as tumor dimensions, texture, and cellular characteristics—to accurately classify breast tumors as malignant or benign. By analyzing clinical measurements derived from mammographic images, the system supports early and precise detection enabling timely medical intervention and improving patient outcomes.

The Proposed System Provides:

A. **High Diagnostic Accuracy** – Utilizes advanced machine learning algorithms to ensure accurate classification of tumor types.

B. **Enhanced Data Privacy** – Ensures secure handling of patient data in compliance with medical data protection standards.

C. **Decision Support** – Assists medical professionals in diagnosis by highlighting high-risk cases and relevant tumor features.

D. **Early Detection Capability** – Promotes proactive treatment throw early identification of malignant early identification of malignant tumors.

3.1 PROBLEM STATEMENT

In the realm of healthcare, there exists a pressing need to develop an accurate and efficient predictive system for the early detection of breast cancer. Breast cancer remains one of the most common and life-threatening diseases affecting women globally. Timely and reliable diagnosis significantly increases the chances of successful treatment and survival. The objective of this project is to harness the power of machine learning techniques to build a robust predictive model capable of distinguishing between malignant and benign tumors based on various diagnostic features. By analyzing tumor characteristics obtained through medical imaging and clinical data, the system aims to assist medical professionals in making informed decisions, reducing diagnostic errors, and improving patient outcomes. Breast cancer is one of the leading causes of death among women worldwide. Detecting it at an early stage can significantly increase the chances of successful treatment and survival. However, the manual diagnosis process through clinical examination, biopsy, and imaging techniques can be time-consuming, expensive, and prone to human error. With the availability of large medical datasets, there is an opportunity to use machine learning algorithms to build models that can predict whether a tumor is benign (non-cancerous) or malignant (cancerous). The main challenge is to create a system that is accurate, reliable, easy to use, and fast so that it can support doctors and patients in the decision-making process.This project addresses the problem by designing a machine learning–based Breast Cancer Prediction System that uses patient diagnostic data and provides automated results through a user-friendly application

3.2 EXISTING SYSTEM

Predicting breast cancer using Python and machine learning relies on existing clinical systems and technologies to enhance early detection and treatment planning.

Computer-Aided Diagnosis (CAD) Systems:

• Assists radiologists by analyzing mammogram images.
• Highlights suspicious areas for further examination.

1. **Electronic Health Records (EHR) Integration**
   • Stores patient medical history, tumor reports, and biopsy results.
   • Enables access to structured data for prediction models.
2. **Tele-oncology Platforms**
   • Remote diagnosis and consultations for breast cancer patients.
   • Enables real-time sharing of imaging and pathology reports.
3. **Clinical Decision Support Systems (CDSS)**
   • Provides risk assessment based on patient history and tumor features.
   • Recommends diagnostic pathways and alerts for follow-up.
4. **Pathology Image Analysis Software**
   • Uses AI and ML to detect cancerous cells in his to pathological images.
   • Improves diagnostic speed and accuracy

3.2 .1 OBJECTIVE

The primary objective of this project is to develop a robust and reliable machine learning-based system . for predicting the likelihood of breast cancer using diagnostic features derived from patient data. By enabling early detection of malignant tumors, the tool aims to facilitate timely medical intervention and improve patient survival rates. It also seeks to evaluate and compare various machine learning algorithms to identify the most effective models for tumor classification. Designed as a non-invasive, data-driven solution, the system reduces reliance on traditional diagnostic procedures while offering healthcare professionals a valuable decision-support tool that enhances diagnostic accuracy and minimizes human error.

- Data Preprocessing – To clean and prepare the breast cancer dataset by handling missing values, removing unnecessary columns, and encoding categorical variables.
- Model Development – To build and train machine learning models (Logistic Regression, SVM, etc.) that can classify tumors into benign and malignant categories
- Performance Evaluation – To test the models on unseen data and measure accuracy, precision, recall, and confusion matrix for reliability.
- Deployment – To develop a simple Flask-based web application where users can input features and get real-time prediction results.
- Support Healthcare – To create a system that can assist medical professionals in early diagnosis and provide a supportive tool for patients

3.3 IMPLIMENTATION:

3.3.1 System Architecture

The system is a web-based machine learning application designed to predict whether a tumor is benign or malignant. Users interact with the app through a browser, where they input medical features in a comma-separated format. These inputs are sent to a Flask backend, which handles the data flow and invokes a preprocessing step using Standard Scaler to normalize the features. Once the data is standardized, it's passed to a pre-trained machine learning model stored as model.pkl. The model analyzes the input and returns a prediction—either 0 for benign or 1 for malignant—which is then displayed back to the user. This architecture ensures a clean, modular flow from user input to intelligent prediction, making it both efficient and scalable.

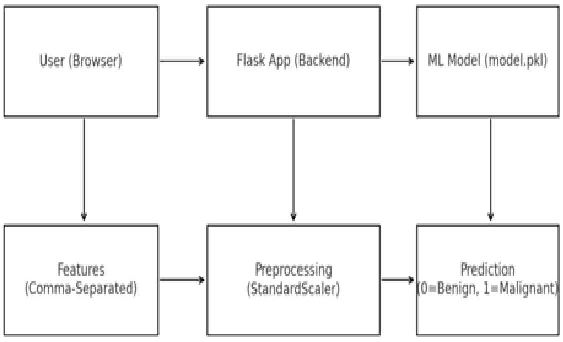

User (Browser)
Flask App (Backend)
ML Model (model.pkl)
Features (Comma-Separated)
Preprocessing (StandardScaler)
Prediction (0=Benign, 1=Malignant)

Figure 1: High-level system architecture from browser to Flask and ML model

3.3.2 Data flow from user input to prediction output

- Dataset: We used the Breast Cancer Wisconsin dataset (diagnosis: M for malignant, B for benign). We removed the empty column 'Unnamed: 32'. We encoded the diagnosis label using Label Encoder so that M/B becomes 1/0.

- Split: We split the data into training (80%) and testing (20%) using train _test _split with a fixed random _state for reproducibility.

- Scaling: We applied Standard Scaler to standardize features to mean 0 and variance 1. This helps models like Logistic Regression and SVM.

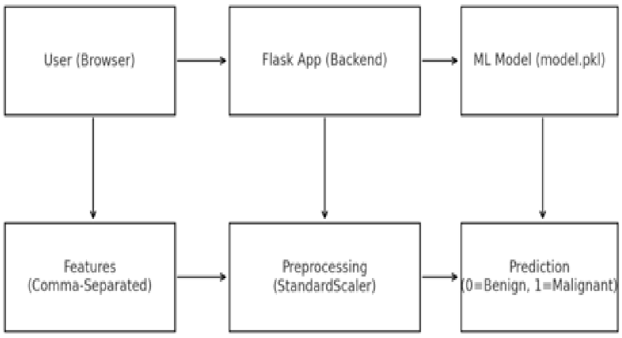
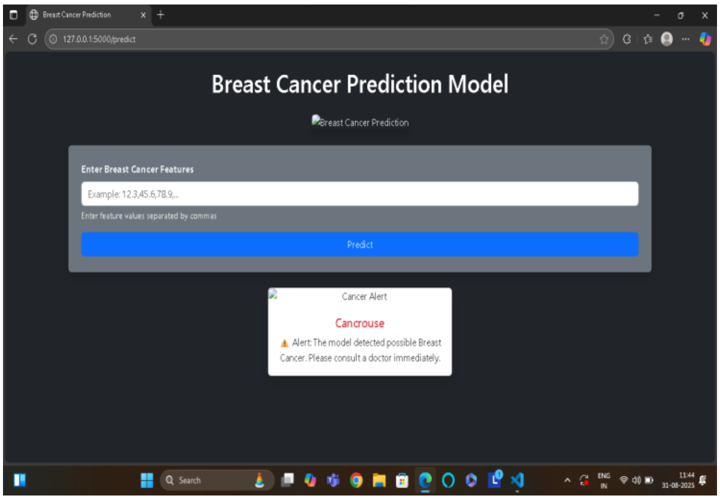

User (Browser)
Flask App (Backend)
ML Model (model.pkl)
Features (Comma-Separated)
Preprocessing (StandardScaler)
Prediction (0=Benign, 1=Malignant)

Figure 2: Data flow from user input to prediction the breast cancer in early stage using ML

3.4 Result



Breast Cancer Prediction Model

Breast Cancer Prediction

Enter Breast Cancer Features

Example: 12.3,45.6,78.9,...

Enter feature values separated by commas

Predict

Cancer Alert

Cancrouse

⚠ Alert: The model detected possible Breast Cancer. Please consult a doctor immediately.

Figure 3: Breast cancer prediction output

3.4.1 Algorithm used for breast cancer detection

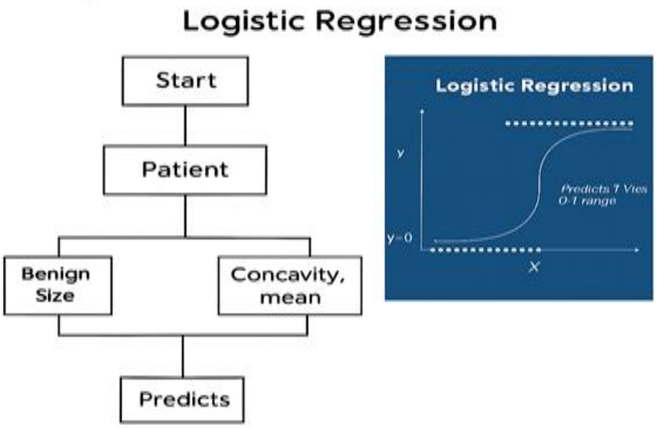


Figure 4:Logistic Regression used for breast cancer detection

In this work, different machine learning techniques were explored for predicting breast cancer, but the primary algorithm used is Logistic Regression.

1. Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for binary classification problems.

In this work, it classifies the tumor as either:

0 → Benign (non-cancerous)

1 → Malignant (cancerous)

It works by fitting the data into a logistic (sigmoid) function that predicts probabilities between 0 and 1.

If the probability is greater than a threshold (usually 0.5), the sample is classified as malignant, otherwise benign.

3.5 CONCLUSION:

The integration of machine learning models for breast cancer prediction marks a significant advancement in diagnostic healthcare. By leveraging clinical and imaging data, these models deliver accurate and early identification of malignant tumors. Machine learning enables personalized risk evaluation, supports medical professionals in informed decision-making, and enhances treatment planning. Through timely detection and data-driven insights, ML-based systems contribute to reducing diagnostic errors, improving survival rates, and ultimately transforming breast cancer care. With continued research and technological collaboration, such intelligent systems hold the potential to save lives and promote better health outcomes for all. Tested the system with real dataset samples to ensure accurate predictions. Verified that the system correctly distinguishes between *Benign* and *Malignant* tumors. Checked accuracy on the test dataset **94%** accuracy simulated development by running the saved model with new unseen data to confirm the stability.

Future Work:

1. Try more models (SVM with tuned kernels, Gradient Boosting).
2. Add calibration for better probability outputs.
3. Add proper forms and validation on the web UI.
4. Deploy the app to a cloud service and secure inputs.
5. Add explainability like SHAP to show which features influenced the prediction.

REFERENCES

[1] .Dua, D., & Graff, C. (2019). “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Dataset”. University of California, Irvine, School of Information and Computer Sciences. Retrieved from https://archive.ics.uci.edu/ml

[2].Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[3].Flask Documentation. (2023). Flask:” Web Development Framework for Python”. Retrieved from https://flask.palletsprojects.com/

[4].Cortes, C., & Vapnik, V. (1995). “Support-vector networks. Machine Learning”, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[5].Sathya, R., & Abraham, A. (2013).” Comparison of supervised learning algorithms for breast cancer diagnosis”. International Journal of Advanced Research in Artificial Intelligence, 2(6), 78–83. https://doi.org/10.14569/IJARAI.2013.020613

[6].Chaurasia, V., & Pal, S. (2017).” Early detection of breast cancer using machine learning” techniques. International Journal of Computer Applications, 162(10), 34–40. https://doi.org/10.5120/ijca2017913211

[7].Ayesha Khan&Farah Shaheen, Springer .(2022).,” Deep Learning Techniques for Breast Cancer Detection”. Shows potential of DL models for image classification in medical imaging, 34-50 . https://ieeexplore.ieee.org/

[8].R. S. Rajput,&P.J. Kulkarni IEEE .(2021).,” Predictive Analytics in Breast Cancer Diagnosis Using Machine Learning”., Helpful for model selection and identifying research gaps. 48-78. https://kalaharijournals.com/

[9].Priya R.& Manish Sharma, Elsevier, (2020).,” A Review of Machine Learning Models for Breast Cancer Detection”. Helpful for model selection and identifying research gaps.,110-127. https://www.irjet.net/

[10].Meenakshi Sundaram,& Kavitha R.,(2019).,” Predictive Analytics in Breast Cancer Diagnosis Using Machine Learning”. Promotes hybrid models as effective alternatives to single classifiers.78-84. https://www.iosrjournals.org/

[11].Md. Sifat Momen & Nargis Akter, (2023).,” Machine Learning Approaches for Breast Cancer Classification” https://pubs.aip.org/

[12].Mohammed Amine Naji a, & Sanaa El Filalib,(2021).,” Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis”., https://ieeexplore.ieee.org/

[13].Kazi Arman Ahmed,& Doulotuzzaman Xames.,(2025)” Advancing breast cancer prediction: Comparative analysis of ML models”., https://www.sciencedirect.com/

[14].Prerita, Nidhi & Alka Chaudhary.,(2021)” Breast Cancer Detection using Machine Learning Algorithms”, https://ieeexplore.ieee.org/document/9596295

[15].Qianqian Guo,& Peng Wu, Junhao He, (2025)” Machine learning algorithms predict breast cancer incidence risk” published in procedia computer science,80-90. https://bmccancer.biomedcentral.com/articles/ 10.1186/s12885-025-14444-