Precision Diagnosis: Leveraging KNN for Breast Cancer Detection

Salka Debbarma[1], Reyad Hossain[2], Niladri Das[2]

[1]Affiliation author 1,
Agartala Tripura,
salka.debbarma98@gmail.com

[2]Affiliation author 2
Agartala Tripura,
reyadhossain16@gmail.com

[2]Affiliation author 2
Agartala Tripura,
niladridas@nielit.gov.in

***Abstract*:**
Breast cancer is a class of disease which is the most common type of cancer nowadays in women and this kind of cancer has millions of new diagnoses globally each year. This research study is focused on early diagnosis for raising cure rates and increasing survival rates among patients. This project will discuss the use of various machine learning models to predict breast cancer, which includes Logistic Regression, Naive Bayes, SVM, K-Nearest Neighbour (KNN), Decision Tree, and Random Forest. KNN outperformed other models with an accuracy of 98.54%, precision of 0.98, and an F1-score of 0.98. To translate this model into a real-world web application, a Flask based web interface was developed to be used by health professionals and patients with real-time predictions. Future work will involve optimization of models, refining features, and integrating the medical system to assist clinical decision-making.

***Keywords*: Breast Cancer Detection, Machine Learning, KNN, SVM, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Flask Web Application**

(Article history: Selected from 3rd NICEDT 2025, Ropar, 14-15 Feb 2025)

I. Introduction

Today, Breast Cancer is the most common tumor diagnosed in women worldwide; therefore, even with improved treatment modalities, mortality rates are still high. Due to the fact that survival rate improves with early detection, conventional methods for the diagnosis of breast cancer, such as mammography and physical examination, are usually associated with a number of problems in terms of accuracy, accessibility, and cost. Consequently, false positive and negative diagnoses are common, which may result in delayed treatment or interventions that are not required.

This study has compared different machine learning models to identify the most accurate model that can be used to the detect breast cancer; after which, the best performing model will be used on a web-based system and made available to healthcare providers who could use it for real time predictions. Traditional diagnostic methods, such as mammography and physical examination often suffer from problems like high costs, low accuracy which leads to false positives or delayed results.

The main goal of this research study is to test the performance of models like Naive Bayes, SVM, KNN, Logistic Regression, Decision Tree, and Random Forest in classifying the disease using a dataset of key diagnostic parameters and Our results describes that KNN outperformed other ML models, achieving the accuracy of 98.54%. We developed a user-friendly web application that integrates the KNN model, which enables healthcare professionals to make real time predictions.

II. Literature survey

A. Survey (1)

In the study" Investigation into the Methods of Breast Cancer Detection," The authors of" Deep Convolutional Neural Network Based on Computer Aid" use Convolutional Neural Networks (CNNs) for image classification and computer-assisted feature extraction. In order to automatically extract characteristic characteristics, they pre-train CNNs with various architectures, and they contrast their technique with conventional methods. With an accuracy of about 89techniques. The goal of the study is to maximize CNN structure's impact on integrating varied data for classification performance [1–3].

B. *Survey (2)*

Mammography is the go-to method for breast cancer screening because it’s one of the best ways to catch cancer early. This technique works by compressing the breast between a device and a compression plate to get a clear image. It uses two main views: one taken from the front and the other from the side. These angles help doctors spot potential issues that could be missed with just one view. [4].

C. *Survey (3)*

In order to improve tumour detection performance at surgical margins with limited labelled data, self-supervised learning is used in the publication” Self Supervised Learning for Detection of Breast Cancer in Surgical Margins with Limited Data”. The model learns patch order to capture model properties, splits picture spectra into smaller batches, shuffles their order to create new instances, and adjusts weights for cancer detection. This method is used on the REIMS dataset, which has 144 cancer data sample sets. Next, the traits that were extracted are used to categorize cancer as benign or malignant [3,5]

D. *Survey (4)*

A. G. N. Sharma,et.al. (2010) discussed about the growth of various type tumour and cancer management of breast cancer [7] In the United States, adenocarcinoma is a frequent kind of cancer in women and the second-leading cause of cancer related deaths among females [6]. The cancer develops in the breast tissue and typically starts in the lining of the milk ducts or the lobules that supply the ducts. Cancer cells have DNA and RNA that are similar to the host organism’s cells, but not identical. Therefore, they are often not recognized by the immune system, especially if the immune system is weakened [6–8]

III. MACHINE LEARNING MODEL

A. Logistic Regression

Logistic Regression is a method in statistics, used to determine the probability in a binary outcome - for example, cancerous versus non-cancerous cells. This model assumes the linearity of features and log-odds of the target variable. It produces the best results in simple classification problems but suffers from its linearity assumption when there is interaction among the features in the form of complex datasets.

B. Naive Bayes

Naive Bayes borrows from Bayes’ Theorem and finds the probability of a data point to a specific class given its features. This method is computation-efficient but, if features are independent of each other (a condition that, though seldom found, still meets satisfactory results), the intrinsic nature of most medical datasets usually shows feature dependencies which, in practice bring down the accuracy.

C. Support Vector Machine (SVM)

SVMs find an optimum hyperplane that separates classes in a feature space. These are highly effective in high-dimensional datasets and thus find broad applications in medical diagnostics. At the same time, SVMs require massive computational resources, especially when applied on large datasets or with complex kernels. In this paper, SVM showed good performance; at the same time, it was slower than KNN.

D. K-Nearest Neighbor (KNN)

KNN is the simplest and most intuitive algorithm that classifies instances of data based on the closest training examples within the feature space. Despite its simplicity, KNN had been able to show the highest accuracy in breast cancer prediction because of the special nature of our dataset. As a non-parametric classifier, it handles the complexity in medical data and does not make assumptions about the distribution of the data.

E. Decision Tree

Decision Trees create a tree-like model of decisions, with nodes representing feature-based splits and leaves representing the class labels. Though interpretable, decision trees tend to overfit, especially on small datasets, and thus do not generalize well on new data.

F. Random Forest

Random Forest generalizes Decision Trees to create a large ensemble of trees, each trained on different subsets of the data. That prevents overfitting and generally increases the accuracy. But this comes at the cost of higher computational complexity; hence, in this case, it is less suitable for realtime applications.

IV. OBJECTIVE

The primary goal of this research aims to enhance the early detection of breast cancer using a prediction model based on machine learning that analyses data and mammograms to make decisions. Using, Flask-based web application, it will enable online predictions, hence reducing false negatives and increasing the improved diagnostics with the K-NN model. The system depends on nine key diagnostic parameters known to offer an accurate and reliable outcome in result by the healthcare professional: Clump Thickness, Uniform Cell Size, Bare Nuclei, etc.

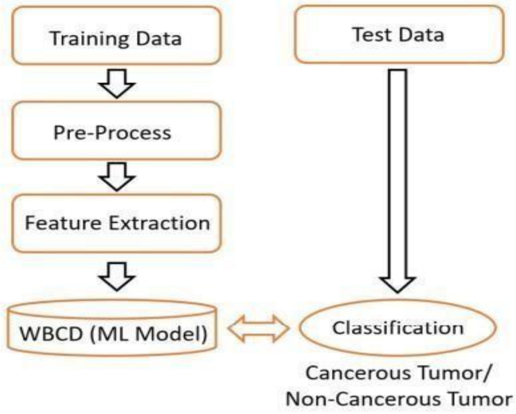


Fig.1. Flow Diagram of work [4].

A. *Data Collection and Pre-Processing*

Before using a dataset for modelling, it is essential to do data pre-processing to raise its quality. To guarantee that the data is correct, consistent, and full, this requires cleaning and changing it, which can eventually improve the effectiveness and dependability of the models based on it [8].

The act of removing unneeded noise and enhancing its overall quality is known as data cleaning. The dataset is often partitioned into two different sets, a training dataset and a validation dataset as part of the data pre-processing step.

- The training dataset is used in training the machine learning model.
- The dataset was divided into 80% for training and 20% for testing to evaluate the model's performance.
- The dataset consists of both benign and malignant cases making it suitable for the binary classification.
- Samples include diagnostic parameters like Clump Thickness, UniSize, UniShape, MargAdh, SingEpiSize, BareNuc, BlandChrom, NormNucl and Mit.
- The validation dataset is utilized during the classification stage [8].

B. *Feature Scaling*

Variables with different units, sizes, and ranges are frequently found in datasets. However, determining the Euclidean distance between data points is a fundamental step in most machine learning algorithms. Scaling techniques must be used to normalise the characteristics so that each one contributes equally to the study. In order to compare the variables on an even playing field, it is necessary to modify the variables to give them comparable scales and ranges [8].

C. *Identify Accurate Model*

Compare multiple ML models such as Logistic Regression, Naive Bayes, SVM, KNN, Decision Trees, and Random Forests for breast cancer diagnosis prediction; Among all the ML models KNN has the highest accuracy (98.54%), precision (0.98), and F1 score (0.98) demonstrating its suitability for breast cancer diagnosis. The second best performing model is Random Forest, achieved an accuracy of 97.81% and precision of 0.97. Logistic Regression and Decision tree performed well but were slightly less accurate at 97.08% and 96.35% respectively.

Table 1. Performance measures of ML models

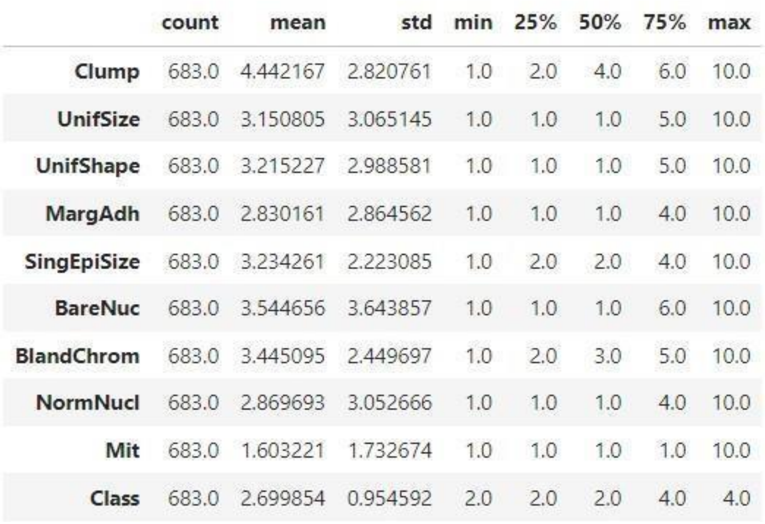
Model	Accuracy	Precision	F1 Score
Logistic Regression	97.08	0.96	0.97

Naive Bayes	94.89	0.94	0.95
SVM-Linear	96.35	0.95	0.96
KNN	98.54	0.98	0.98
Decision Tree	96.35	0.96	0.96
Random Forest	97.81	0.97	0.98

ML Model information is given in Table 1.

D. *Evaluate key diagnostic parameters*

After selecting the model with the best accuracy, which is KNN ML model, we calculate the 9 specific parameters from the dataset to detect whether a patient has cancer. These parameters include Single Epithelial Cell Size, Uniform Cell Shape, Bland Chromatin, Marginal Adhesion, Clump Thickness, Normal Nucleoli, Mitoses (mit), Bare Nuclei



	count	mean	std	min	25%	50%	75%	max
Clump	683.0	4.442167	2.820761	1.0	2.0	4.0	6.0	10.0
UnifSize	683.0	3.150805	3.065145	1.0	1.0	1.0	5.0	10.0
UnifShape	683.0	3.215227	2.988581	1.0	1.0	1.0	5.0	10.0
MargAdh	683.0	2.830161	2.864562	1.0	1.0	1.0	4.0	10.0
SingEpiSize	683.0	3.234261	2.223085	1.0	2.0	2.0	4.0	10.0
BareNuc	683.0	3.544656	3.643857	1.0	1.0	1.0	6.0	10.0
BlandChrom	683.0	3.445095	2.449697	1.0	2.0	3.0	5.0	10.0
NormNucl	683.0	2.869693	3.052666	1.0	1.0	1.0	4.0	10.0
Mit	683.0	1.603221	1.732674	1.0	1.0	1.0	1.0	10.0
Class	683.0	2.699854	0.954592	2.0	2.0	2.0	4.0	4.0

Fig.2. These parameters are used to detect Breast Cancer Using KNN ML Model.

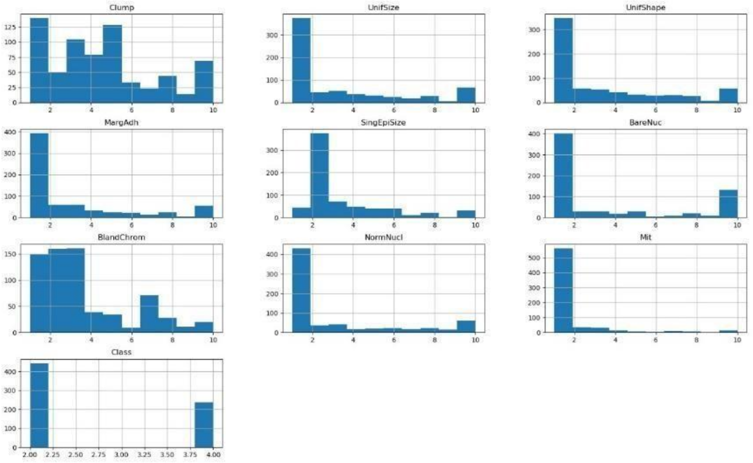
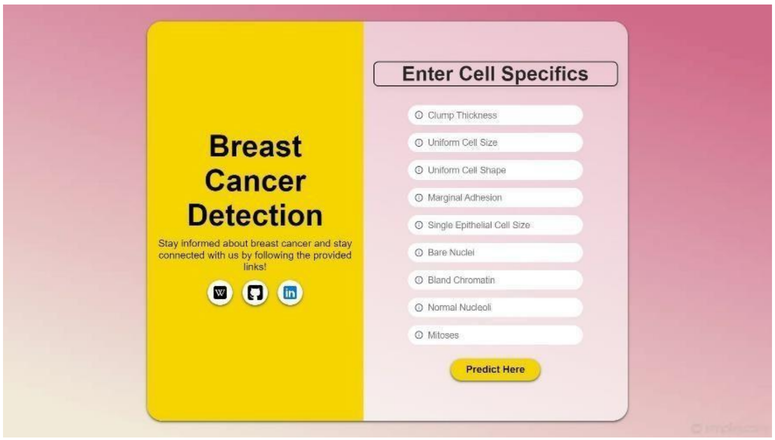


Fig.3. Data visualization of parameters which is helpful to detect cancer.

E. *Web-Based Application Using Flask*

We build a web application from scratch using Flask that make predictions on breast cancer based upon the best-performing model in real-time, which is KNN: After that, we calculate the parameters like Single Epithelial Cell Size, Uniform Cell Shape, Bland Chromatin, Marginal Adhesion, Clump Thickness, Normal Nucleoli, Mitoses (mit), Bare Nuclei from the dataset to detect whether a patient has cancer or not. And the parameters like Single Epithelial Cell Size, Uniform Cell Shape, Bland Chromatin, Marginal Adhesion, Clump Thickness, Normal Nucleoli, Mitoses (mit), Bare Nuclei are required to detect cancer cells.



Breast Cancer Detection

Stay informed about breast cancer and stay connected with us by following the provided links!

Enter Cell Specifics

Fig.4. Web-based app interface to detect Breast cancer.

F. *Portability*

And, we are also making it accessible for doctors as well as for patients; for diagnosis of breast cancer using the machine learning web-based application.

V. CONCLUSION AND FUTURE DIRECTION

A. *Conclusion*

Among the compared models, the KNN model proved to be the best for the breast cancer detection cases with accuracy of 98.54%. Because of its simplicity and superior performance, the K-NN model was ideal to be implemented for practical use. Indeed, we have really implemented this solution, thus making it available for any user via web application with a backend created in Flask. The web application will allow a user to input all the patient data relevant for this problem and get a prediction in real time.

B. *Future Work*

1. **Model Tuning:** Further tuning will be done on the KNN model, and techniques of hyper parameter optimization such as cross-validation and grid search will be attempted to improve the accuracy in the future.
2. **Feature Enhancement:** The other variables that may further improve the performance of this model include family history, genetic markers, and lifestyle factors.
3. **Deep Learning Integration:** Deep learning models, especially CNNs, can be attempted for medical imaging data analysis together with structured patient data for better prediction.
4. **UI Extension:** Future development will involve user authentication, role based access, and rich data visualization in order to enhance user experience.
5. **Integration with Medical:** Web application integration with EHR systems in hospitals will offer health professionals predictive insights during patient consultation.

REFERENCES

[1] Yogesh Suresh Deshmukh, Parmalik Kumar, Rajneesh Karan, and Sandeep K Singh.Breast cancer detection-based feature optimization using firefly algorithm and ensemble classifier. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 1048–1054. IEEE, 2021.

[2] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[3] Dipali Ghadge, Shrutik Hon, Tushar Saraf, Tejas Wagh, Abhishek Tambe, andYS Deshmukh. Analysis on machine learning-based early breast cancer detection. In *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, pages 1–5. IEEE, 2024.

[4] Madhuri Gupta and Bharat Gupta. A comparative study of breast cancer diagnosisusing supervised machine learning techniques. In *2018 second international conference on computing methodologies and communication (ICCMC)*, pages 997–1002. IEEE, 2018.

[5] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, NatashaAntropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89– 94, 2020.

[6] Kavita Saini, Ishika Mishra, and Shreya Srivastava. Farmer's e-mart: An e-commercestore for crops. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 346–350. IEEE, 2021.

[7] Ganesh N Sharma, Rahul Dave, Jyotsana Sanadya, Piush Sharma, and KK22247839 Sharma. Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology & research*, 1(2):109–126, 2010.

[8] Pratyaksh Singh, Jaideep Nagill, and Kavita Saini. Using supervised learning forbreast cancer detection using ai&ml. In *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 281–285. IEEE, 2023.