# Breast Cancer Detection Using ML


| | |
|---|---|
| Kritika Khandelwal | Vidyavardhinis College of Engineering and Technology |
| Atul Mishra | Vidyavardhinis College of Engineering and Technology |
| Shreyash Seth | Vidyavardhinis College of Engineering and Technology |
| Dr. Swapna Borde | Vidyavardhinis College of Engineering and Technology |



**Abstract**

*Breast cancer is the second most commonly diagnosed cancer among women globally, and early detection is crucial to improving survival rates. Machine learning (ML) and artificial intelligence (AI) techniques have shown promising results in detecting breast cancer in medical imaging data. In this study, we explore the application of ML and AI algorithms for breast cancer detection using mammogram images.*

*Breast cancer detection is a critical healthcare challenge worldwide, and Machine Learning (ML) and Artificial Intelligence (AI) are increasingly being used to aid in the detection process. ML algorithms can analyse vast amounts of data, identify patterns, and classify tumours with high accuracy.*

*The main idea here is to utilize all the open source datasets and breast cancer detection methodologies such as K-nearest neighbour, Convolutional Neural Network, Support Vector Machines, Generative Adversarial Networks to identify pros and cons of all the methodologies. The result of this would be to find the most efficient model to work in a particular scenario.*

*Moreover, machine learning algorithms can be trained to predict breast cancer risk and personalise screening recommendations for individual patients. As such, AI and ML have enormous potential in the fight against breast cancer, improving the diagnosis and treatment of the disease.*
*There are several machine learning algorithms available that are used in this system including KNN, SVM, CNN, GANS, Decision Tree, Random Forest, K-means.*


*Keywords—Decision Tree, Random Forest, Convolution Neural Networks, Support Vector Machine, K-Nearest Neighbour, Machine Learning, Breast Cancer Detection.*

## I INTRODUCTION

Breast cancer is one of the top causes of death among women. But, early detection of cancer helps in preventing it. If breast cancer is diagnosed early, the chances of survival are very good. Breast cancer is a disease that arises, but when a woman or man notices this symptom, it quickly progresses beyond its first stage. Breast cancer is a common and severe disease in women. Cancer is the development of aberrant cells that are genetically and altered. Different techniques are used to capture breast cancer such as Ultrasound Sonography, Computerised Thermography, Biopsy (Histological images). Machine Learning techniques and algorithms are a straightforward way to understand the data and predict it.

The radiologist examines and analyses himself, and then decides on the outcome after consulting with other professionals. This process takes time, and the results are dependent on the staff's knowledge and experience. Furthermore, experts are not available in every field around the world. Therefore, the research community proposed an automatic system called CAD (Computer-Aided Diagnosis) for better classification of tumours, which helps in accurate results and faster implementation without the need for radiologists or specialists.

According to the most recent cancer statistics, breast cancer accounts for 25% of all new cancer diagnoses and 15% of cancer deaths in women globally. In the event of any indication or symptom, people usually see a doctor right away, who may advise you to an oncologist for assistance. An oncologist can diagnose breast cancer by thoroughly reviewing the patient's medical history, inspecting both breasts, and evaluating for swelling or hardening of any lymph nodes in the armpits. This process is lengthy and takes time to detect cancer, which is an important factor for the person diagnosed with cancer. Thus, the main problem arises to detect breast cancer in patients efficiently to save their time and the efforts by engaging themselves in various processes.

The accuracy of CART model was 91% whereas Naïve Bayes model was 92% accurate. The accuracy of SVM model was more as compared to KNN model which was 94% accurate.

## II. LITERATURE REVIEW

Clustering methods such as the K-nearest algorithm and Support Vector Machines have been used to diagnose breast cancer using metrics such as the radius mean, smoothness mean, concavity mean, and symmetry mean of the breast.

Below is the table of comparative study of the IEEE papers which we have referred to regarding this detection system.

| Sr. No | Authors | Focus of the Paper | Key points in the coverage | Techniques Used | Parameters/Datasets | Research Gaps/Limitations |
|---|---|---|---|---|---|---|
| 1. | Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad | Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks | Histology images, convolutional neural networks, Breast cancer detection | CNN, Neural Networks, Feature extraction | Histology images, the growth rate of cancer cells and tissues. | The performance of the multi-class classification is low and can be improved. |
| 2. | Kundan Kumar, Annavarapu Chandra Sekhara Rao | Breast Cancer Classification of Image using Convolutional Neural Network | Breast Cancer, Convolutional Neural Network, Image Classification, Deep Learning | Multilayer perceptron, Rectified Linear Unit, CNN | BreakHis dataset consists of 9,109 pictures of breast tumor tissue. | Classificationion accuracy mainly depends on how CNN extracts and learns the feature in different layers |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3. | Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan | Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors | Prediction; Support Vector Machine; K-Nearest Neighbors; Performance Measure Indices. | SVM, KNN, Performance Measure Indices | Breast Cancer Wisconsin Data Set | The performance of the KNN method is less and can be improved with the larger dataset. |
| 4. | Saira Charan, Muhammad Jaleed Khan, Khurram Khurshid | Breast Cancer Detection in Mammograms using Convolutional Neural Network | CNN, Mammograms, X-rays | Preprocessing, Convolution Neural Network, Feature Extraction | Mammograms MIAS dataset consisting of histology images | no proper segmentation for efficient feature extraction and classification |
| 5. | Abdul Qayyum, III. Basit | Automatic Breast Segmentation and Cancer Detection via SVM in Mammograms | CAD, Breast cancer, Mammograms, Mini-MIAS database, Otsus segmentation, GLCM. | Otsus segmentationon, median filtering, Support Vector Machines | mini-MIAS dataset | The model's accuracy is less than AlQoud which a larger dataset can increase. |
| 6. | Naresh Khuriwal, Dr. Nidhi Mishra | Breast Cancer Detection From Histopathological Images Using Deep Learning | Deep Learning,CNN, Neural Network, Random Forest, Support Vector Machine, Machine Learning, MIAS Dataset | Data Pre-processsing, Convolution Neural Network, Feature Extraction | MIAS Dataset from Histopathology images containing 200 images | This paper worked on only 12 features. Real images dataset can be used for best result and accuracy. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7. | Madhuri Gupta, Bharat Gupta | A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques | Machine Learning; Classification ; Linear regression; Decision Tree; Multilayer perceptron; Breast Cancer | KNN, SVM, Decision Tree, Multilayer Perceptron | Wisconsin Prognostic Breast Cancer (WPBC) | In order to further improve accuracy, Info Gain test, Gain Ratio test will be incorporated in future. |
| 8. | Abdul Qayyum, III. Basit | Automatic Breast Segmentation and Cancer Detection via SVM in Mammogram | CAD, Breast cancer, Mammograms, Mini-MIAS database | Otsus segmentation, median filtering, Support Vector Machines | mini-MIAS dataset | The accuracy of the model is less than AlQoud which can be increased by larger dataset. |

Table1: Comparative Study of IEEE Paper

## III. PROPOSED ALGORITHM

Kaggle provided us with the Breast Cancer Wisconsin (Diagnostic) Dataset. In this case, 569 patients' knowledge was analysed; each instance contains 32 attributes with identification and options, but the id parameter is not used in training or testing. Every instance contains a parameter of carcinogenic and non-cancerous cells, and we may predict cancer simply by entering options. The table below shows the dataset attributes.

| Dataset | No. of Attributes | No. of Instances | No. of Classes |
|---|---|---|---|
| Wisconsin Breast Cancer (Original) | 11 | 699 | 2 |
| Wisconsin Diagnosis Breast Cancer (WDBC) | 32 | 569 | 2 |
| Wisconsin Prognosis Breast Cancer (WPBC) | 34 | 198 | 2 |

Table2: Dataset Attributes

Data modulation was performed on the Wisconsin dataset which is a process of transforming the original dataset to better suit the needs of the machine learning algorithm. In the case of the Wisconsin dataset, which contains data on breast cancer, data modulation techniques are used to preprocess the dataset before feeding it to the machine learning model. This involved steps such as feature scaling, normalisation, and handling missing values. Feature scaling was used to ensure that all the features have a similar range of values, so that one feature does not dominate the others. Normalisation was used to transform the data into a standard format, so that different features can be compared more easily.

Handling missing values was done through imputation techniques such as mean imputation or median imputation, where the missing values are replaced with a value derived from the other values in the same feature. By using data modulation techniques, the Wisconsin dataset was transformed into a format that is better suited for machine learning algorithms, thereby improving the accuracy and effectiveness of the model.

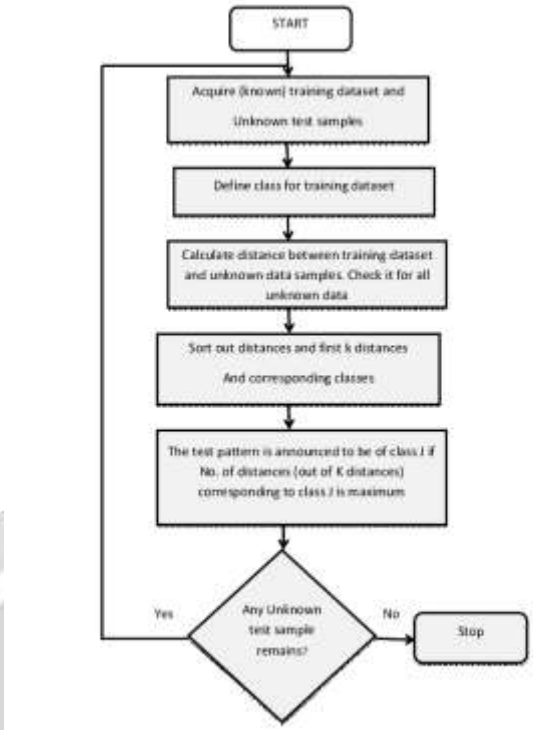


Figure1: Flow chart of Breast Cancer Recognition System using KNN.

Initially training a model using Naive Bayes or CART on the Wisconsin dataset requires data modelling and in the next step splitting the data into training and testing sets and training the model on the training set using the selected algorithm. Data splitting is performed using the exhaustive or the non-exhaustive methods. The parameters of the algorithm would be optimised to find the best combination of hyperparameters that yield the highest accuracy. Finally, performance of the trained model is evaluated on the testing set by measuring its accuracy.

Further, Support Vector Machines (SVM) and K-Nearest Neighbours (KNN) algorithms were trained on the Wisconsin dataset. SVM is a powerful algorithm that tries to find the best separating hyperplane between two classes of data. It maximises the margin between the classes and can handle non-linear data with the help of a kernel function.

KNN is a simple yet effective algorithm that classifies new data points based on the k closest training examples. It requires no training time and can be used for both classification and regression tasks.

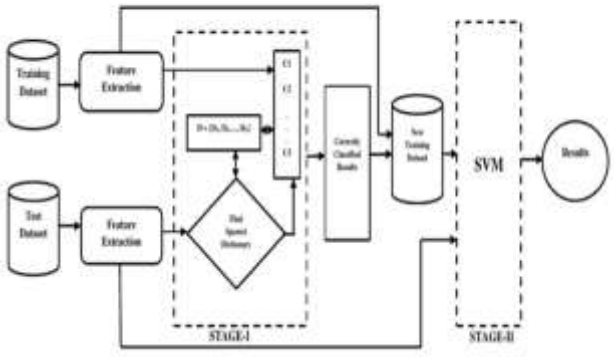


Figure2: Architectural diagram of Breast Cancer Recognition System using SVM.

To train a model using SVM or KNN on the Wisconsin dataset, pre-processing the dataset using techniques such as feature scaling and normalisation was the first step. The data would then be divided into training and testing sets, and the model would be trained on the training set using the chosen algorithm. The parameters of the algorithm would be optimised using techniques such as grid search or random search to find the best combination of hyperparameters that yield the highest accuracy.

K-NEAREST NEIGHBOUR:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. The working of KNN is demonstrated with the diagram as follows:

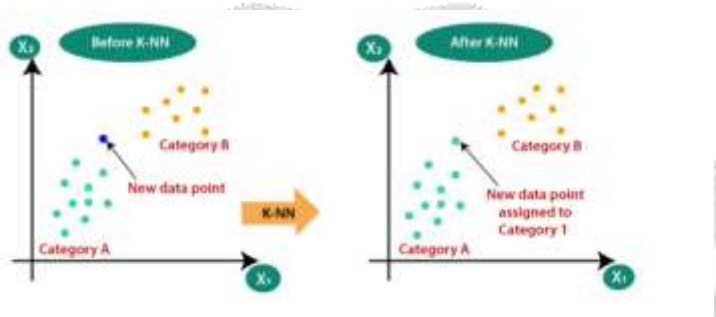


Figure3: Working of SVM

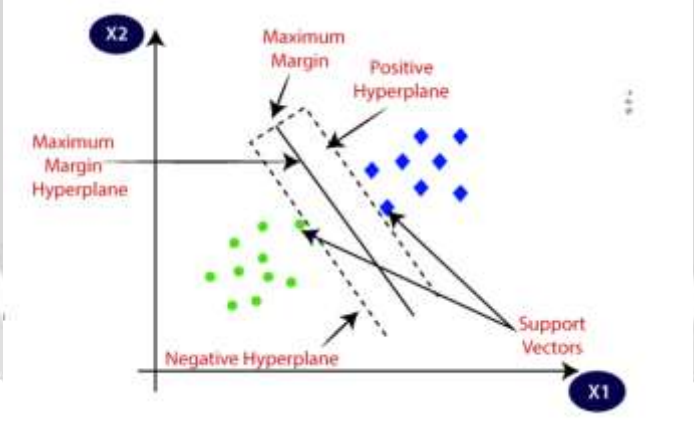


Figure4: Working of SVM

SUPPORT VECTOR MACHINES:

SVM, or the Support Vector Machine, is a supervised algorithm for machine learning which can be applied to both classification as well as regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

The accuracy of the training model would then be measured in order to assess its performance on the testing set. The optimum algorithm for our particular application can be chosen by contrasting the performance of SVM and KNN on the Wisconsin dataset. These formulas are employed to determine whether a cancer is benign or malignant.

## IV. RESULTS

The exploratory Data Analysis can be explained in terms of countplot of Malign and Benign Cancer.

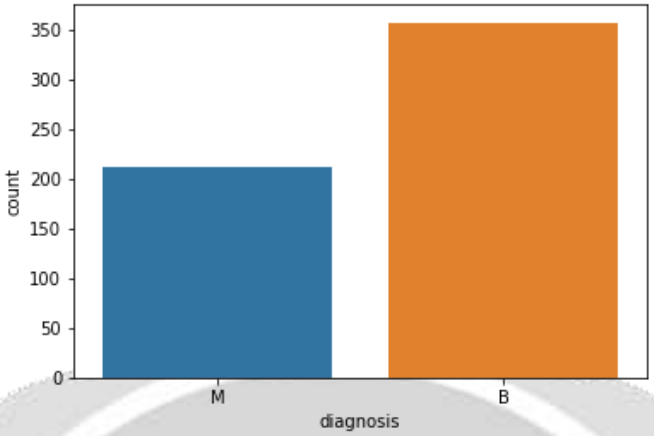


Figure5: Count plot of malignant and benign

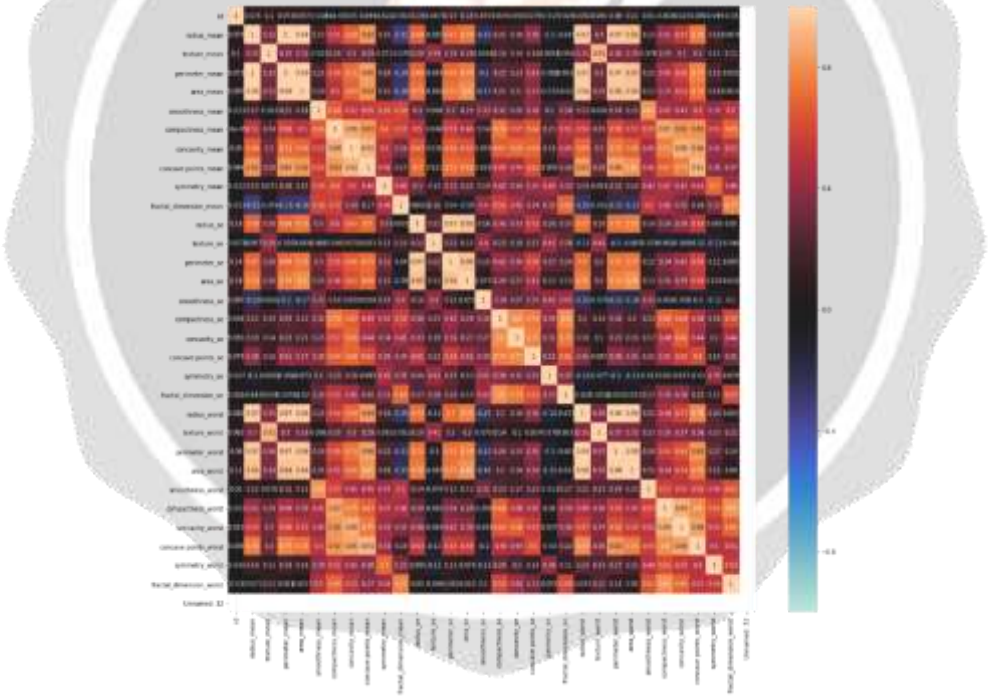

Figure6: Correlation Matrix

The Confusion matrix is created after training the model on SVM.

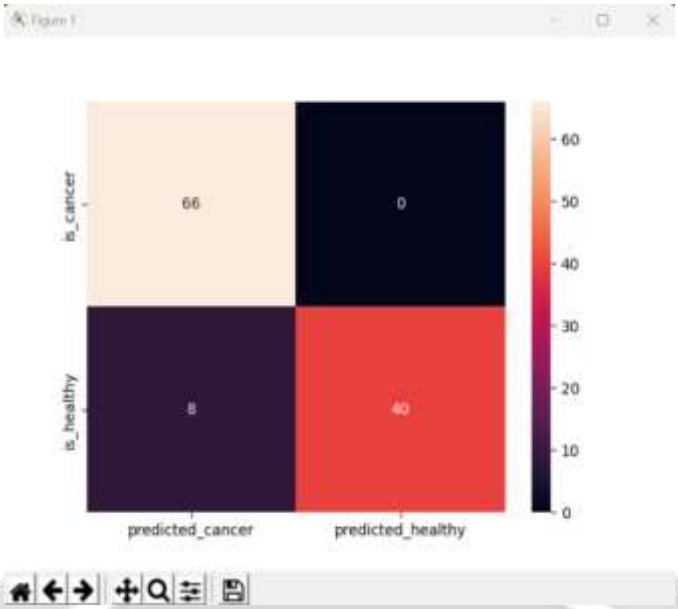


Figure7: Confusion Matrix

The SVM model delivers greater performance and accuracy in comparison to the KNN model. The SVM model performed better, with a mean accuracy of 90% on the training data. On the training data, the KNN model had an accuracy of 80%.

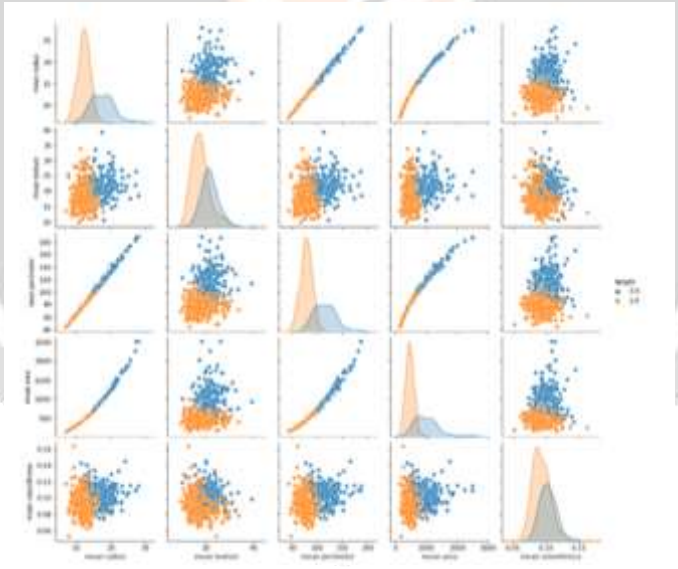

Figure8: Data Visualization

## V. CONCLUSION

After pre-processing and filtering the data, we trained our model on the Wisconsin dataset using multiple models, beginning with a basic logistic regression classifier. In order to have a better understanding of the entire dataset and the various parameters that can be utilised to train the model, we did data visualisation on it. The original and the diagnostic versions of the Wisconsin dataset, each containing 10 and 32 features, were employed.

In the holdout approach, the data is divided into training and testing portions in the ratios of 80:20 or 70:30. To determine which of the two is more effective for the detection and prediction processes, we used the k-fold cross validation approach in the exhaustive method.

Following the Logistic Regression model, we extensively trained the SVM and KNN models and tested them on both datasets using both exhaustive and non-exhaustive Cross Validation.

Hence, the proposed work helps in easy and fast detection of breast cancer, helping those in need and saving the life of many women.

## VI. REFERENCES


1. DeSantis C, Siegel R, Bandit P, Jamail A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods." 2018 Electric
2. H. Wang.: Nearest Neighbours without k: A Classification Formalism based on Probability, technical report, Faculty of Informatics, University of Ulster, N.Ireland, UK (2002)
3. G. Guo1, H. Wang, D. Bella, Y. B, and K. Geer, "KNN Model-Based Approach in Classification," Lecture Notes in Computer Science, pp 986-996, 2003.
4. Y.-S. Sun0, Z. Zhao S5, Z.-N. Yang3, F. Xu3, H.-J. Lu5, Z.-Y. Zhu4, W. Shi1, J. Jiang0, P.-
5. P. Yao0, and H.-P. Zhu8, ''Risk factors and preventions of breast cancer,'' Int. J. Biol. Sci., vol. 13, no. 11, p. 1387, 2017.
6. Reddy, V. Anji0, and Badal Soni6. "Breast Cancer Identification and Diagnosis Techniques." Machine Learning for Intelligent Decision Science. Springer, Singapore, 2020. 49-70.
7. Y. Lu9, J.-Y.9 Li, Y.-T. Su9, and A.-A. Liu9, ''A review of breast cancer detection in medical images,'' in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2018, pp. 1–4.
8. Dalal Bardou0, Kunn Zhang, and Sayed Mohamad Ahmad3 et. al, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks" vol. 10, p. 187, 2019.
9. Abdul Qayyumm, Av. Basit, "Automatic Breast Segmentation and Cancer Detection via SVM in Mammograms" with mini MIAS data
10. Manav Mangukiyaa, Anooj Vaghani, Meet Saavani, "Breast Cancer Detection with Machine Learning",
11. Link: https://towardsdatascience.com/
12. Link: https://hindawi.com/