Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7641457

# BREAST CANCER DETECTION SYSTEM

**J.Grace Priyanka**
Assistant Professor, Department of Computer Science and Engineering, Dr.G.U.Pope College of Engineering, Tuticorin.

**P. Daniel Raj**
Assistant Professor, Department of Computer Science and Engineering, Dr.G.U.Pope College of Engineering, Tuticorin.

## Abstract

Breast cancer is the leading cause of cancer death. India has witnessed 30% of the cases of breast cancer during the last few years and it is likely to increase. Breast cancer in India accounts that one woman is diagnosed every two minutes and every nine minutes, one woman dies. The chances of correct treatment and survival are greatly increased by early diagnostics, but this process is tedious and often leads to a disagreement between pathologists. Early detection and diagnosis can save the lives of cancer patients. Issues such as technical reasons, which are related to imaging quality and human error have increased the misdiagnosis of breast cancer in radiologists' interpretation. In the effort to overcome such restrictions, CAD systems are developed to automated breast cancer detections and classify benign and malignant lesions. Computer-aided diagnosis systems have the potential to improve diagnostic accuracy. By A Computer Aided Diagnosis system, Breast cancer can be detected as early as possible. By Early prevention the chances of death can be reduced . Our Project presents a method to detect breast cancer by employing techniques of Machine Learning. The carried out an experimental analysis on a dataset to evaluate the performance. The proposed method has produced highly accurate and efficient results when compared to the existing methods. This Project utilizes the CNN algorithms for the High Accuracy in the results and the prediction of cancer Affected percentage .Overall , this project seek to the early detection of breast cancer by Hypothetical images with high accuracy using CNN and overcome the drawbacks in the existing Systems.

## Introduction

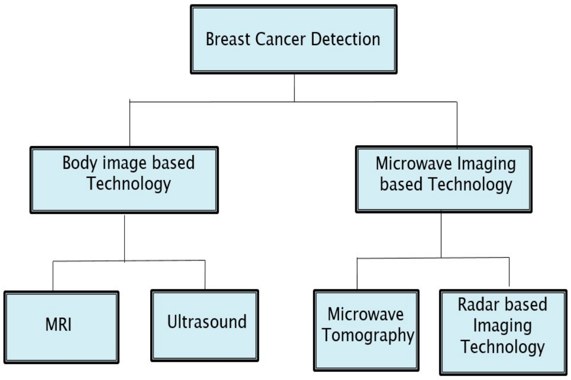
Breast cancer (BC) is the malignant tumor that activates in the cells of the breast. A tumor has the potential to spread to other parts of the body [2]. BC is a universal disease that hammers the lives of women typically in the age group of 25- 50. With the potential rise in the number of BC cases in India, the distress reaching in alarming proportions. During the past five years, the survival rates of BC patients are about 90% in the USA and whereas in India the figure reports approximately 60% [3]. BC projection for India during 2020 suggests the number to go as high as two millions [4]. Specialist Doctors have identified hormonal, way of life and environmental factors that may increase an individual's odds of developing BC. Over 5-6% of BC patients have linked to gene mutations that went through the ages of the family. Obesity, increasing age, postmenopausal hormonal imbalances are the other factors that cause BC. As such, there is no prevention mechanism for BC, but early detection can significantly improve the outcome. In the effort to overcome such restrictions, CAD systems are developed to automated breast cancer detections and classify benign and malignant lesions. The CAD systems improve radiologists' performance in finding and discriminating between the normal and abnormal tissues. These procedures are performed only as a double reader but the absolute decisions are still made by the radiologist. Recent advances in the resolution of medical imaging modality have revolutionized the diagnostic accuracy. Efficient usage of imaging data to improve the diagnosis is very important. In recent years, computer-aided diagnosis systems (CADs) have developed a new context in radiology to take advantage of the data to be applied to different imaging modality and the diagnosis of different diseases. The efficiency of radiologists' interpretation can be improved in terms of accuracy and consistency in detection or diagnosis while productivity can be improved by reducing the time required to read the images [6]. The outcomes are derived using various techniques in computer vision to present some of the significant parameters such as the location of suspicious lesions and the likelihood of malignancy of the detected lesions. In relation to breast cancer, the main objective of CAD system is to design accurate and reliable approach, to decrease observational oversights and assist in discriminating benign and malignant lesions. In the following, we present some of the recent proposed CAD systems for breast cancer detection or diagnosis on biopsy histopathological images.

Hispothatical images is a dedicated imaging modality for breast screening that uses low dose X-ray during breast examination. It is currently the most effective tool for early detection of breast cancer; however, it has some restrictions. Breast density is a variety of confounding factors that make diagnosis of breast cancer more difficult in women with dense breasts (Ertosun and Rubin, 2015). The contrast between cancer and background in dense breast image is very low, which can affect the diagnosis outcome (Longo et al., 2014). In the mammographic examination, non-cancerous lesions can be misinterpreted as cancer (false-positive value), while cancers may be missed (false-negative value). As a result, radiologists fail to detect 10 % to 30 % of breast cancers [7,8].The false-positive value indicates the percentage of lesions that are found to be cancerous and subjected to biopsy. The miss rate in mammography has increased in dense breasts where the probability of cancer is four to six times higher than in non-dense breasts[7,9,10]. Several solutions have been proposed to enhance the specificity and sensitivity of mammography as well as to decrease unnecessary biopsies procedure. Double reading is one of the solutions that can significantly contribute to achieving high sensitivity and specificity [11,12]. Additional costs will be imposed on the patients for double reading of mammography. CAD systems can be considered as an alternative framework that acts as a second reader to enhance the performance of physician's interpretation. The studies [13,14 ,15,16] have shown that the attention to use a computer to improve the performance of physicians to detect mass and micro-calcification in mammography has increased in recent years. [17] indicated that proportion of cancer detected was 199 of 227 (87.7 %) for double reading and 198 of 227 (87.2 %) for single reading with CAD system. The perspective assessment of the impact of CAD systems on interpretation mammogram images has been performed on a community of breast cancer patients[18].Among 12,860 mammograms, the radiologist's performance was measured without CAD and with CAD. The recall rate increased from 6.6 % to 7.7 % and the proportion of early stage malignancy detected the growth from 73

% to 78 %, which represents an increase in efficiency in the detection of cancer with the usage of CAD system. Further this can also considerably reduce the costs of the treatment. However, sometimes it is unusual to show cancer symptoms, so early detection is difficult. It's indispensable to employ mammograms and self-breast tests to detect any early irregularities before the tumor gets advanced [5]. The key objective of this paper is to propose a novice method to detect BC. This paper presents a detailed study of existing cancer detection models and presents the highly accurate and efficient results.

## Related Works

In this session, we introduce a number of past studies related to the Breast Cancer Detection System. Further, authors reviewed various Data sets from regional and national cancer registries. Different modalities such as mammography, ultrasound, and Magnetic Resonance Imaging (MRI) are the most effective tools in the early detection of breast cancer.


Breast Cancer Detection
Body image based Technology
Microwave Imaging based Technology
MRI
Ultrasound
Microwave Tomography
Radar based Imaging Technology

The authors have taken most popular BC detection methods namely; Naïve Bayes Classifier, Support Vector Machine (SVM) Classifier, Bi-clustering and Ada boost Techniques. These methods are described in this section.

SVM Classifier technique [19] is an amalgamation of RFE and SVM.

RFE is a technique that operates by choosing dataset features depending on the least feature value in a recursive manner.

Accordingly, SVM-RFE is operated by removing the inappropriate features (lowest weight feature) in all iterations.

AdaBoost is a most renowned ensemble technique and it is proficient of enhancing the accurateness of classification by combining several weak classifiers. The bi-cluster oriented classifiers can also be integrated with a strong ensemble classifier for superior generalization performance. During training, diverse weights are allocated and decisions are made depending on "weighted majority voting".

## System Analysis

Analysis is the first crucial step, Analysis is defining the boundaries of the system that will be followed by design and implementation. This section gives the detailed study of the various operations performed by a system and their relationships within and outside of the system.

## Existing Algorithms

Existing Machine Learning Models namely:
Naïve Bayes (NB),
Support Vector Machine(SVM),
Decision Tree (DT),
Bi-clustering and
Ada boost Techniques .

**Naive Bayes**

Naive Bayes classification algorithms based on Bayes' theorem which is powerful to the predicted variables. Naïve Bayes algorithm is to is classified the group of data items to efficient and, correct, fast. It is more accept in different group of data prediction analysis. When we assume of non-dependence data variables is handle, a Naive Bayes algorithm to perform the good compare to other model like regression analysis. It is good performing give to the different input data compare the numeric value of variable in data, for numeric value of distribution is predicted.

**SVM**

SVM classification algorithm to classify customers according to their buying behaviour. Classification is done by considering how the customer spends their valuable time, day in buying decisions.

**Decision Tree**

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

**Bi-Clustering**

Biclustering algorithm simultaneously cluster rows and columns of a data matrix. These clusters of rows and columns are known as biclusters. Each determines a submatrix of the original data matrix with properties.

**AdaBoost Technique**

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an Ensemble Method. The most common estimator used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees are also called Decision Stumps.

**Disadvantages**

- Naive Bayes Classifier produces ill results when the training data is not represented [21].
- The SVM classifier is unsuitable for large data sets and also not effective on high computer vision applications.
- When the data is imbalanced, Biclustering and Ada boost Techniques will lead to erroneous classification.
- RCNN takes more time to train the network. HA-BiRNN may produce wrong scores for BUS images [20].

**Proposed System**

In Our System, we propose a breast cancer histopathology image classification by Deep learning model of the patients hispothatical image through a web page.

First, the datasets of the cancer affected patients Hispothatical images where collected.
The collected data sets where tested and trained by Deep Machine Learning Techniques. The Algorithm used in this detection is Convolutional Neural Network (CNN) one of the popular technique in Deep learning.CNN is the most promising technique for detecting cancer in hispothatical images with the accuracy.

A brief description of the main stages of a CAD system is provided as follows:

**(1) Image pre-processing**:

This step is essential for some modality such as ultrasound for the purpose of enhancing the image and reducing the noise with minimum distortion of image features. Some of the CAD systems do not have a pre-processing stage.

**(2)Image segmentation**:

Image segmentation is a vital step towards efficient development of CAD systems. The main purpose of segmentation is the separation of the region of interest (ROI) commensurate with the desired properties [24]. Recently, imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), 3D ultrasound, and many more modalities are capable of producing images in the form of 3D. Therefore, 3D segmentation methods are desirable for more accurate segmentation in volumetric imagery.

**(3)Feature extraction and selection**:

In this step, different features are extracted according to the characteristics of lesions from the image. These features are used to distinguish benign or malignant lesions. The feature set is usually very large and the selection of the most effective features is very critical for the next step.

**(4) Classification:**

According to the selected features, the suspicious areas are classified to benign or malignant based on different classification techniques. The common classification methods used in medical imaging are presented in this section.

**(5) Performance evaluation:**

This step evaluates the performance of CAD system. Image pre-processing In medical image processing, image preprocessing plays a significant role to achieve the ideal outcomes in other stages of a CAD system such as segmentation and feature extraction. Pre-processing stage is performed to remove noise and defect caused in image acquisition procedure, image resizing, and enhance the image intensity [25].

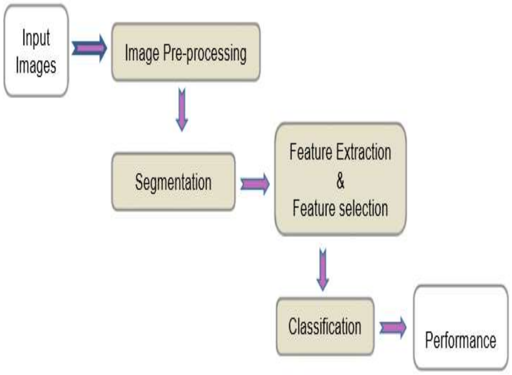
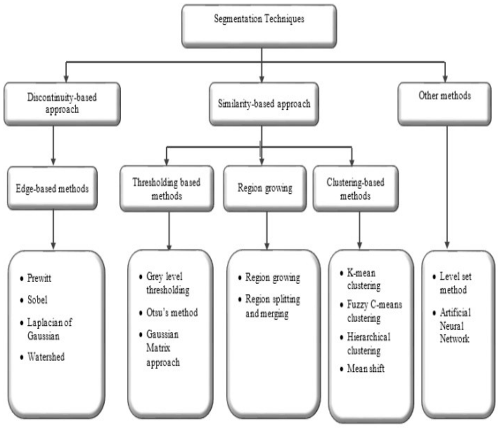


**Image segmentation**

Image segmentation is a crucial component in computer vision and pattern recognition. Accurate segmentation plays a significant role in the success or failure of the computerized procedure. In medical imaging aspect, the selection of segmentation methods are widely depending on the specific application and imaging modality. With increased dimension and resolution of the image in various modalities, the images cannot be manually examined with regard to the huge amount of image information. Segmentation techniques help to highlight significant regions and extract various structures such as organs or tumors for further examination. Segmentation methods are categorized generally into two groups: semi-automatic and fully automatic. Providing an automatic algorithm in medicine to detect and localize abnormality is highly desirable. In medical application, the low rate of false positive and false negative detection is very important. Therefore, evaluation methods of segmentation algorithms is another dimension of CAD systems in the practical aspect. Segmentation approaches based on image properties are broadly arranged into two groups: discontinuity-based approach and similarity-based approach. Discontinuity based approach partitions an image based on an abrupt change in intensity [28]while similarity based method partitions an image according to pre-determined similarity criteria. The similarity-based method is categorised into region-based, thresholding-based, and clustering-based methods. A general comparison of segmentation methods (Lee et al., 2015; Narkhede, 2013) are provided in Table 1.

Edge-based segmentation methods Edge-based segmentation methods are a structural technique to detect edges or pixels among different regions that have abrupt intensity change. The edge based method works well on high contrast and non-noise images. There are several methods for edge-based segmentation such as Sobel, Prewitt, Laplace, Canny, and Laplacian of Gaussian. The main application of edge-based segmentation techniques is human organ recognition. A mathematical morphological edge detection algorithm has been presented in to detect lungs in CT images that contain salt-and-pepper noise. Haris et al. have proposed an integration of edge-based and region-growing with watershed transform for 2D/3D segmentation of magnetic resonance images. Thresholding-based segmentation methods One of the wide methods used for image segmentation is a thresholding-based technique which is an effective way to discriminate foreground from the background image [26]. The first step in this method is the selection of an appropriate threshold value according to image properties, and then the pixels image is assigned to specific regions. The automatic selection of threshold value requires the knowledge on the intensity characteristics of the objects, sizes of the objects, and the number of various types of objects existing in the image[27]. The thresholding-based methods have been widely used to develop CAD systems in order to extract significant areas for additional analysis. In the article of [29] various thresholding techniques have been compared to segmentation mammogram images. An automatic nucleus segmentation is developed on the image of breast histopathology using histogram-based thresholding. The result shows 97 % accuracy in nucleus detection.


Segmentation Techniques
Discontinuity-based approach
Similarity-based approach
Other methods
Edge-based methods
Thresholding based methods
Region growing
Clustering-based methods
• Prewitt
• Sobel
• Laplacian of Gaussian
• Watershed
• Grey level thresholding
• Otsu's method
• Gaussian Matrix approach
• Region growing
• Region splitting and merging
• K-mean clustering
• Fuzzy C-means clustering
• Hierarchical clustering
• Mean shift
• Level set method
• Artificial Neural Network

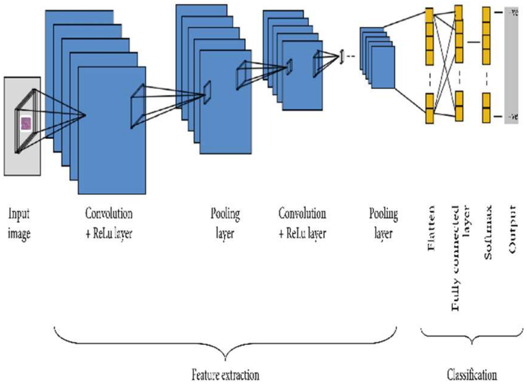
## Feature extraction

Computing feature descriptors from an image to reduce the volume of data ordinarily signifies feature extraction. Features are characteristics of the whole image or region of interests. The proper selection of features has an important influence on (1) memory size, (2) accuracy of classification, (3) computational cost of classification and (4) robustness. Feature descriptors and metrics widely depend on the specific application. Generally, image descriptors are divided into the three dimensions. Based on feature descriptor dimensions using three axes: shape, pattern and spectra, and density. According to the literature image descriptors are placed in three categories, namely shape based, textural, and colour-based descriptor,. Shape features are the important properties employed by human to discriminate objects with other features such as colour and texture. To address the complexity conversion of shapes, an effective shape descriptor should be invariant into the rotation and scaling.

## CNN

CNNs are applied to explore patterns in an image. This is done by convoluting over an image and looking for patterns [23]. The network can detect lines and corners in the few front layers of CNNs. Via our neural net, however, we can then transfer these patterns down and begin to identify more complex characteristics as we get deeper. This property ensures that CNNs are very effective at detecting objects in images [22]. The proposed system uses CNNs to detect breast cancer from breast tissue images.

The architecture of a CNN has 3 main layers, the convolutional layer, pooling layer, and fully connected layer . The first layer calculates the output of neurons which are linked with local regions. Each one is calculated by a dot product of weights and the region. For image inputs, typical filters are small in area such as 3 × 3, 5 × 5, or 8 × 8. These filters scan the image by a sliding window on the image, while learning the recurrent patterns which arise in any area of the image. The interval between filters is known as the stride. The convolution is extended to overlapping windows if the stride hyper parameter is smaller than the filter dimension. A detailed visual explanation of neural networks (NNs).



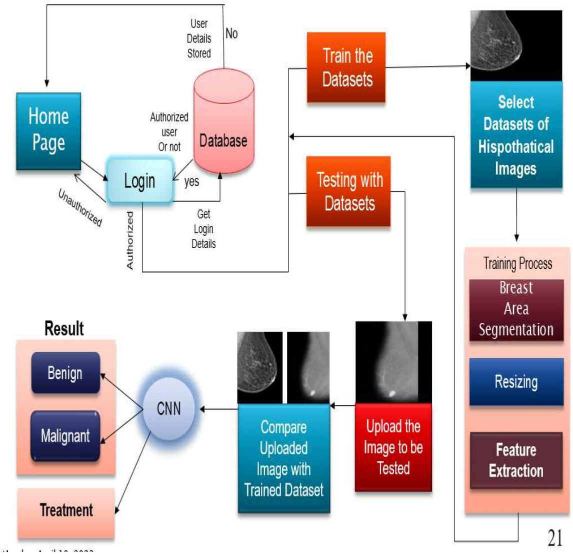
**Advantage**

- Accurate Prediction of Breast Cancer.
- No Risk about Radiations.
- Takes much less Time to detect.
- Cost Efficient when compare with existing systems
- Easy to detect earlier.

**System Architecture**

**1.Breast Cancer Detection System using Hispothatical images.**

In this project we developed the system for Detect the Breast Cancer using the Hispothatical images by train and test the images with the High accuracy along with the treatment suggestion.



## 2. Data Set Annotation

In Training Phase “We used Hispothatical images from Kaggle ,the dataset, Kaggle 162 H&E, was used for the proposed system . Kaggle 162 H&E was also used by many researchers for similar kind of study.The data set consists of both benign and malignant images. Careful observation was ensured during splitting; the dataset was divided into validation data and testing data belonging to same distribution to well represent the model’s generalized results. For learning indicators like weights and biases, training data is important, while validating data is essential for model verification and how exactly the model simplifies, thus tuning hyper parameters like learning rate and decay to boost the result of the model. A model’s final output comes from precise work on the test results. To hold each pixel in the same range and prevent bias, the normalization has to be done on the whole image. Around 277,524 50 × 50-pixel RGB digital image patches were extracted from 162 WSI mounts scanned samples.

In Testing Phase “In this project, user login train and test the dataset, and give accurate result about the affected percentage of breast cancer.

## 3.Pre-processing

The main goal of the pre-processing is to improve the image quality to make it ready to further processing by removing or reducing the unrelated and surplus parts in
the background of the mammogram images Mammograms are medical images that complicated to interpret. Hence pre-processing is essential to improve the quality. It will prepare the mammogram for the next two-process segmentation and feature extraction. The noise and high frequency components removed by filters.

A.Mean filter or average filter

The goal of the mean filters used to improve the
image quality for human viewers. In this, filter replaced each pixel with the average value of the intensities in the neighbourhood. It locally reduced the variance, and easy to carry out. Limitations of average filter I)

Averaging operations lead to the blurring of an image, blurring affect features localization. II) If the averaging operations applied to an image corrupted by impulse noise, the impulse noise attenuated and diffused but not removed. III) A single pixel with a very unrepresentative value affected the mean value of all the pixels in neighbourhood significantly.

B.Median filtering

A median filter is a nonlinear filter is efficient in removing salt and pepper noise median tends to keep the sharpness of image edges while removing noise. The several of median filter is I) Centre-weighted median filter II) weighted median filter III) Max-median filter, the effect of the size of the window increases in median filtering noise removed effectively.

C.Adaptive median filter

Adaptive median filter works on a rectangular region Sxy. It changes the size of Sxy during the filtering operation depending on certain conditions as listed below. Each output pixel contains the median value in the 3-by-3 neighbourhood around the corresponding pixel in the input images. Zeros however, replace the edges of the images. The output of the filter is a single value, which replaces the current pixel value at (x, y), the point on which S is centered at the time. The following notation is used:
Zmin = minimum pixel value in Sxy
Zmax = maximum pixel value in Sxy
Zmed = median pixel value in Sxy
Zxy= pixel value at coordinates (x, y)
Smax = maximum allowed size of Sxy
Adaptive Median filtering used to smooth the non-repulsive noise from two-dimensional signals without blurring edges and preserved images. This makes, it particularly suitable for enhancing mammogram images. The preprocessing techniques used in mammogram, orientation, label, artifact removal, enhancement and segmentations. The preprocessing involved in creating masks for pixels with highest intensity, to reduce resolutions and to segment the breast.

D.Wiener filter

The Wiener filter tries to build an optimal estimate of the original image by enforcing a minimum mean square error constraint between estimate and original image. The wiener filter is an optimum filter. The objective of a wiener filter is to minimize the mean square error. A wiener filter has the capability of handling both the degradation function as well as noise. From the degradation model, the error between the input signal f(m, n) and the estimated signal f(m, n) is given by

$$E(M,N)=F(M,N)-F(M,N) \qquad (1)$$

The square error is given by

$$[F(M,N)-F(M,N)] \qquad (2)$$

The mean square error is given by
E{[F(M,N)-F(M,N)]}                    (3)

## 4.Feature Extraction

Convolutional Neural Network (CNN) is a special kind of deep neural network model.

### Two-dimensional spatial relationship of pixels.

We define the gray level co-occurrence matrix with direction $\phi$ and interval d as $P(i,j,d,\phi)$. Among them, L is the number of gray levels of the image, and the data in this study is 16-bit data. The co-occurrence matrix is a huge matrix, which makes it difficult to calculate the post-processing. Therefore, it is necessary to compress the gray level of the data without affecting the image texture information. In this paper, we use the AUC value of single feature under different gray levels (G = 8, 16, 32, 64, 128, 256) to judge. The AUC value is an important indicator for evaluating the classification effect.

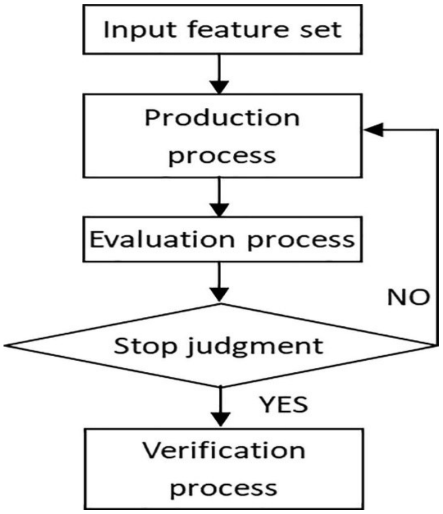
### Calculation of gray level co-occurrence matrix

The meaning of $P(i,j,d,\phi)$ is the probability that the starting pixel has a gray value of i and the ending gray level is j when the distance is d and the direction is $\phi$. specific case of the calculation of the gray level co-occurrence matrix, is a 5 × 5 original image. According to statistics, the gray value of the original image has three values, namely 0, 1, and 2, so the gray level co-occurrence matrix has a dimension of 3 × 3. Assuming that the direction is $\phi = 135°$ and the distance is $d = \sqrt{2}$, the matrix is finally obtained by statistics.

### Adjacent relationship of voxels in three-dimensional space.

Traditional texture analysis focuses on two-dimensional images or two-dimensional projection images. For 3D images, it is common practice to select the most representative slices and approximate the three-dimensional results using the analysis results of the two-dimensional slices. The downside is that extracting information from 2D slices is not ideal and ignores vertical information between slices. With the development of 3D medical images, 2D texture analysis methods can no longer satisfy the research on 3D images. Therefore, this study optimizes the 2D texture feature extraction method and uses 3D texture feature extraction method to study the image. The two-dimensional gray level co-occurrence matrix is expanded. The pixels adjacent to the center pixel in two dimensions have eight points, and there are 26 voxels extending into the three-dimensional image adjacent to the central voxel, and the shape .In a two-dimensional plane, only one angle $\phi$ is needed for direction control, and when extended to three dimensions, two directions $\phi$ and $\theta$ need to be introduced.

**Process of feature selection**



## 5.Classification

Classification is the last stage in CAD systems that differentiates and labels of the abnormality. Classification methods play an important role in the diagnosis and educational purposes in medicine. Classification approaches are categorized into two groups. Generally, in the implementation of classifier in clinical image processing, supervised classification techniques are used. Supervised classification examines a large number of unknown data and assigns them into related classes based on their characteristics. The main difference between unsupervised and supervised methods is that the unsupervised do not require pre-determined class. In a successfully supervised classification, all classes should be defined and the spectral properties of these classes have to be extracted during the training phase. However, in unsupervised classification, classes may be discovered but not known in advance. A brief description of the most popular supervised classification techniques along with their advantages and disadvantages are proposed .Impact of imbalanced data set in classification The imbalance dataset is a crucial issue in various pattern recognition applications. In binary classification, this problem occurs when the number of instances from one class is significantly less than the other class. In this situation, the overall predictive accuracy is achieved by the majority class while the minority class has a greater impact on the classifier performance. The impact of the imbalanced data in the real-world applications is the irreversible effect on classification performance, specifically in medical diagnosis. Due to delays in diagnosis and treatment, the patient may lose their lives. To deal with the imbalanced dataset, several approaches have been presented in functional level and data level .Kernel transformation techniques and biased penalties approaches are recommended schemes for boosting support vector machines in functional level .

## Results and Discussion

We have used scikit-learn machine learning framework for implementation in Python. Scikit-learn is most popular among data scientists. Also, there are other prerequisites to run scikit-learn functions such as pandas, NumPy, matplotlib, and seaborn frameworks which have been used to implement the proposed system.

**Predicting Invasive Cancer Using Machine Learning Classifiers**.

It presents the accuracies of machine learning (ML) classifiers. The highest level of accuracy is found in SVMs when compared to logistic regression (LR) and k-NN.

**Predicting Invasive Cancer Using CNN Model 1**.

CNN Model 1 has two convolution layers with 32 and 64 kernels, which are checked with a dropout regularization of 25% to cancel over fitting. The image is then vectored with a flattened layer for the next dense layer. The rectified linear unit is an activation function that is used in all layers with the exception of the output layer, for which the Soft max activation function is used. This model has been trained with 12 epochs and the batch size is 128. The training loss is 0.69, while the validation is 0.68. Little difference is found between model's performances in the training set and the validation set. Table 2 shows the configuration summary of Model 1 with the metrics results presented in Figure 8. It has 59% accuracy, which is less than standard machine learning (ML) classifiers. **Predicting Invasive Cancer Using CNN Model 2.**

To increase the number of features, convolution layers are tripled here. The accuracy of the proposed system is thus increased to 0.76, an improvement on Model 1.The confusion matrix and the loss learning curve are the validation score is consistently less than the training score, with the suspicion that this model suffers from bias.

**Predicting Invasive Cancer Using CNN Model 3.**

CNN Model 3 is deeper than Models 1 and 2, with a five-layer CNN used to detect breast cancer,is architecture gives the best result with 87% accuracy: it also provides a similar distribution of predicted labels to that of actual labels (50/50).

## Conclusion

Automating the detection of breast cancer to enhance the care of patients is a challenging task. , Our Project proposes a CNN approach that detects breast cancer using the hispothatical images with the high accuracy rate. As of late, computer aided design frameworks are created to mechanize bosom disease identifications and arrangement of harmless and threatening sores in various modalities like ultrasound, mammography, and X-ray. The computer aided design frameworks work on radiologists' presentation in finding and segregating among ordinary and unusual tissues. The principal phases of execution of computer aided design framework and various procedures for every particular step were classified and introduced in this section. The locale based division and bunching based calculations are ridiculously used to foster computer aided design frameworks for bosom disease location. Remove appropriate elements for the recognition of ordinary and unusual sores in bosom relies upon the idea of mass and imaging modalities, in which different highlights were presented in this part. Counterfeit canny strategies and backing vector machines have been generally examined to foster grouping outline work in the determination of bosom disease as of late. The unevenness dataset is a vital issue in different example acknowledgment applications. To manage the imbalanced dataset a few methodologies have been introduced in utilitarian level and information level .Part change procedures and one-sided punishments approaches are suggested plans for helping support vector machines in useful level. At the information level, over-examining and under-inspecting procedures are broadly used to defeat imbalanced dataset issues. Engineered Minority Oversampling Procedure and Versatile Manufactured Testing Approach (ADASYN) are successful oversampling methods which have some lack, for example, over-age in light of the fact that the age of engineered tests expands the classes covering.

## Future Enhancement

In the future, we will apply our approach in other similar applications and evaluate its performance by using other datasets. In conclusion, The result is also affected by the quality and the number of images used. On the other hand, with the current technological advances, any technological findings will be able to be improved by the next technological findings. Likewise, the detection of breast cancer through CNN-XGBoost on Mammogram Images also allows the accuracy to be improved in future.

## References

[1] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, Omar Elgendy "Breast cancer detection using AI: A systematic literature review" 02276, 2022(IEEE).

[2] Hotko YS. Male breast cancer: clinical presentation, diagnosis, treatment. Exp Oncol 2013; 35:303-10

[3]https://www.biospectrumindia.com/views/21/15300/statistical-analysisof-breast-cancer-in-india.html

[4] Malvia S, Bagadi SA, Dubey US, Saxena S.Epidemiology of breast cancer in Indian women, Asia Pac J Clin Oncol. 2017 Aug;13(4):289-295

[5] Shallu, Rajesh Mehra, Breast cancer histology images classification: Training from scratch or transfer learning?, ICT Express 4 (2018), 247– 254 Sixun Ouyang; AonghusLawlor , "Improving Explainable Recommendations by Deep Review-Based Explanations" , 2021,[Online].Available:https://ieeexplore.ieee.org/document/9417205

[6] Doi K. Computer-aided diagnosis in medical imaging: achievements and challenges. Paper presented at the World Congress on Medical Physics and Biomedical Engineering, Munich, Germany, 2009.

[7] Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007;356:227-36.

[8] Bird RE, Wallace T W, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology. 1992;184:613-7.

[9] Maskarinec G, Pagano I, Chen Z, Nagata C, Gram IT. Ethnic and geographic differences in mammographic density and their association with breast cancer incidence. Breast Cancer Res Treat. 2007;104:47-56.

[10] Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L, et al. Screening for breast cancer: an update for the US Preventive Services Task Force. Ann Intern Med. 2009;151:727-37;W237-42.

[11] Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. Breast. 2001;10:455-63.

[12] Warren R, Duffy W. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Brit J Radiol. 1995;68 (813):958-62.

[13] Kerlikowske K, Carney PA, Geller B, Mandelson MT, Taplin SH, Malvin K, et al. Performance of screening mammography among women with and without a firstdegree relative with breast cancer. Ann Intern Med. 2000;133:855-63.

[14] Sanchez Gómez S, Torres Tabanera M, Vega Bolivar A, Sainz Miranda M, Baroja Mazo A, Ruiz Diaz M, et al. Impact of a CAD system in a screen-film mammography screening program: A prospective study. Eur J Radiol. 2011;80:e317-21.

[15] Malich A, Azhari T, Böhm T, Fleck M, Kaiser W. Reproducibility - an important factor determining the quality of computer aided detection (CAD) systems. Eur J Radiol. 2000;36:170-4.

[16] Marx C, Malich A, Facius M, Grebenstein U, Sauner D, Pfleiderer SOR, et al. Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of mammographic diagnosis with and without use of CAD. Eur J Radiol. 2004;51:66-72.

[17] Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med. 2008;359:1675-84.

[18] Freer TW, Ulissey M J. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. Radiology. 2001;220:781-6.

[19] Anji Reddy.V, Badal Soni, Breast Cancer Identification and Diagnosis Techniques, Machine Learning for Intelligent Decision Making, Springer 2020.

[20] Huang, Qinghua, Yongdong Chen, Longzhong Liu, Dacheng Tao, and Xuelong Li. On Combining Biclustering Mining And Adaboost For Breast Tumor Classification, IEEE Transactions on Knowledge and Data Engineering, Vol. 32, Issue 4 (2020) 728-738.

[21] Kharya, Shweta, and SunitaSoni. Weighted naive bayes classifier: A predictive model for breast cancer detection, International Journal of Computer Applications 133, no. 9 (2016) 32-37.

[22] P. Kumar, S. Srivastava, R. K. Mishra, and Y. P. Sai, "End-to-end improved convolutional neural network model for breast cancer detection using mammographic data," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, Article ID 154851292097326, 2020.

[23] M. Alhussein, G. Muhammad, M. S. Hossain, and S. U. Amin, "Cognitive IoT-cloud integration for smart healthcare: case study for epileptic seizure detection and monitoring," *Mobile Networks and Applications*, vol. 23, no. 6, pp. 1624–1635, 2018.

[24] Nie K. Development of breast MRI computer-aided diagnosis system. Thesis. Irvine, CA: University of California, 2009.

[25] Kyaw MM. Pre-segmentation for the computer aided diagnosis system. Int J Computer Sci Inf Technol. 2013:5(1):79.

[26] Zhang Y-J. An overview of image and video segmentation in the last 40 years. Adv Image Video Segment. 2006;1-15.

[27] Al-Amri SS, Kalyankar NV. Image segmentation by using threshold techniques. J Comp. 2010;2(5):83-6.

[28] Rastgarpour M, Shanbehzadeh J. Application of AI techniques in medical image segmentation and novel categorization of available methods. Paper presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists 2011, Vol I, IMECS 2011, March 16-18, 2011, Hong Kong.