Breast Cancer detection using Machine Learning Algorithms

Vibha S [1] , Shailaja K.P [2], Pushpa T.S [3]

[1] Dept. of Computer Applications, B.M.S. College of Engineering, Bangalore, India.
[2] Dept. of Computer Applications, B.M.S. College of Engineering, Bangalore, India.
[3] Dept. of Computer Applications, B.M.S. College of Engineering, Bangalore, India.

E-mail: [1]vibhas.mca21@bmsce.ac.in, [2]shailaja.mca@bmsce.ac.in, [3]pushpa.mca@bmsce.ac.in

***Abstract*— The most frequent cancer in women globally is breast cancer, and effective treatment of the condition depends on early identification. Machine learning, a branch of artificial intelligence, has seen growing use in recent years in breast cancer detection and diagnosis using medical imaging. The use of machine learning algorithms in the cancer field can potentially improve the accuracy and efficiency of breast cancer detection. There are several machine learning techniques that have been applied in breast cancer detection, including supervised learning, deep learning, machine learning and unsupervised learning. The evaluation can be done using metrics such as sensitivity, specificity, accuracy, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The algorithm learns to identify patterns and features in the images that are indicative of cancer.**

***Keywords*—** Breast Cancer, Support Vector Machines (SVMs),Random Forests,K-mediods.

I. INTRODUCTION

Machine learning has emerged as a promising method for aiding in the early diagnosis of breast cancer in recent years. Breast cancer is a leading cause of mortality among women globally.

Machine learning algorithms analyse data from mammograms, medical records, and other sources to identify patterns and relationships that might indicate the presence of cancer. This information can be used to develop a predictive model that can assist radiologists in identifying potential cases of breast cancer. The models can also be used to highlight areas of the mammogram that require further examination, reducing the chances of missed diagnoses.

The ability of machine learning to handle enormous volumes of data and make predictions based on patterns and relationships in the data is one advantage of utilising it in the detection of breast cancer. This may result in more precise and reliable results compared to traditional methods that rely on manual analysis by radiologists. Additionally, machine learning models can be continually trained on new data, allowing for continuous improvement in their accuracy over time.

However, it is important to note that machine learning is not a replacement for traditional diagnostic methods, but rather an aid for radiologists. The results generated by machine learning algorithms should be interpreted by a qualified medical professional to make a final diagnosis.

In order to increase patient survival rates, machine learning has the potential to revolutionise the early identification of breast cancer.

II.ARCHITECTURE OF CANCER PREDICTION STSTEM

Unsupervised learning methods, such as k-means and hierarchical clustering, have been used to identify patterns and features in breast tissue that may indicate the presence of cancer. These algorithms work by analyzing unlabeled data, where the inputs are not paired with any outputs. The algorithm identifies patterns in the data and groups similar inputs together.

Convolutional neural networks (CNNs), a type of deep learning technique, have also been used to analyze mammography, magnetic resonance imaging (MRI), and ultrasound pictures to find breast cancer. An artificial neural network design called a CNN is ideally suited to image analysis. They operate by breaking out photos into layers, with each layer identifying more intricate elements.

According to recent studies, machine learning can increase the precision with which malignant tumors are detected when used to diagnose breast cancer.

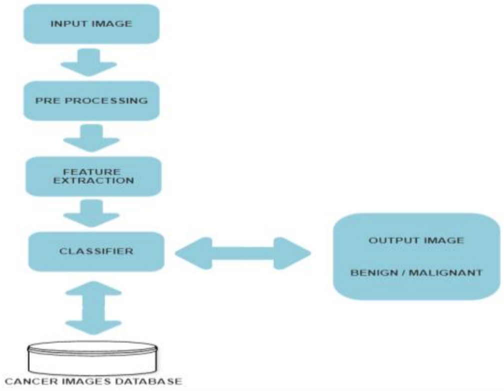


Figure 1: Architecture of Cancer Prediction System

Cancer detection is a complex and challenging task that requires a sophisticated architecture to accurately diagnose and identify the presence of the disease. The architecture for cancer detection typically consists of several stages, each designed to address a specific aspect of the problem.

The first stage is image pre-processing, which involves the cleaning and normalization of medical images. This stage is crucial in ensuring that the images are suitable for analysis and interpretation.

The next stage is feature extraction, where features such as shape, texture, and intensity are extracted from the images and used to train a classifier. The classifier is then used to make predictions about the presence of cancer.

In the final stage, a decision module is employed to evaluate the predictions made by the classifier. This stage includes the use of statistical analysis, machine learning algorithms, and data visualization techniques to make a final diagnosis.

The architecture for cancer detection also includes a mechanism for continuous improvement, where the results of previous predictions are used to improve the accuracy of future predictions. This can be achieved through ongoing research and development efforts aimed at refining the existing algorithms and developing new ones.

III. WORKING OF THE MODEL

K-Medoids algorithm involves grouping mammogram images into two or more clusters based on their similarity, with the idea that each cluster represents either benign or malignant tissue. The algorithm can be used in conjunction with other machine learning techniques to improve the accuracy of the diagnosis. The process of detecting breast cancer using K-Medoids can be broken down into the following steps:

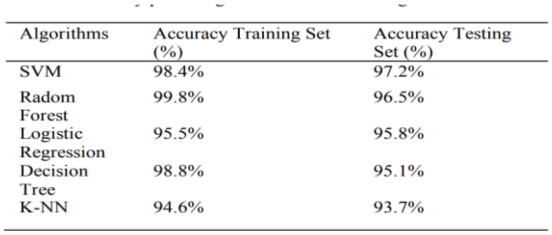
1.Data collection and preprocessing: This involves collecting a large dataset of mammogram images and preprocessing the data by normalizing the intensities and removing any irrelevant information.

2.Feature extraction: This involves extracting relevant features from the mammogram images that can differentiate between benign and malignant cases. This can involve techniques such as thresholding, segmentation, and texture analysis.

3.Clustering: The extracted features are then used to perform K-Medoids clustering, which groups the images into two or more clusters based on their similarity. Unlike K-Means, K-Medoids uses a representative medoid from each cluster rather than a mean, which can result in a more robust clustering.

4.Diagnosis: The cluster assignments can then be used to diagnose the mammogram images as either benign or malignant.

5.Evaluation: Finally, it is important to measure the performance of the algorithm using measures like accuracy, precision, recall, and F1 score to gauge performance and make any necessary adjustments.



Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.4%	97.2%
Radom Forest	99.8%	96.5%
Logistic Regression	95.5%	95.8%
Decision Tree	98.8%	95.1%
K-NN	94.6%	93.7%

Figure 2: Comparison of various Algorithms

a) Image pre-processing:

Image preprocessing in artificial intelligence refers to the manipulation of raw image data to improve its quality, enhance its features, and make it suitable for further analysis and processing by machine learning algorithms. This can involve operations such as noise reduction, contrast enhancement, normalization, resizing, and image segmentation, among others.

The Gaussian filter, which offers the images smoothness, can be used for resizing the input photos. In order to improve the facial image, illumination can be decreased, and disparity can be minimized with the median filter by using normalization as a preprocessing technique.

b) Noise removal

Noise removal is an important step in pre-processing data for Machine Learning. The presence of noise in data can negatively impact the performance of ML models, leading to incorrect predictions or results.

There are several techniques for removing noise from data, including filtering methods and dimensionality reduction techniques. Filtering methods involve applying mathematical operations to the data to remove or reduce the impact of noise. For example, a Gaussian filter can be applied to an image to smooth out small variations in pixel intensity and reduce the impact of noise.

Noise can enter the digital image during digitization and transmission during image acquisition. In the majority of real-time scenarios, imaging sensors are susceptible to interferences from the environment, which causes noise to be added to images as they are transmitted.

Let the actual image be represented by a (xi, yi), the noise by (xi, yi), and the output pixel by r (xi, yi).

$r\ (x_i, y_i) = a\ (x_i, y_i) + \eta\ (x_i, y_i)$

We can reconstruct the image if we can estimate the noise. An alternative nonlinear method for removing noise from photos is median filtering. It is quite good at shielding the edges and at reducing noise. It is a very effective approach to get rid of "salt and pepper" noise. Every pixel in the image is examined by the median of its nearby pixels as part of the median filtering process. A "window" is a pixel and its neighbors that scrolls across every other pixel in the full image.

A window's median is determined by placing all of its pixels in ascending order, then replacing the pixel under consideration with the use of median pixel.

c) Image Segmentation:

A picture is segmented by splitting it into separate portions or segments, each of which corresponds to a different object or part of an object that is visible in the image. It is a crucial phase in computer vision and is used for many different things, such as object detection, image compression, and medical image analysis.

In Machine Learning, Image Segmentation is often tackled using a supervised learning methods, where the model is trained on a dataset of annotated images and learns to predict the segmentation masks for new images. Convolutional Neural Networks (CNNs) have shown excellent performance in this task and are widely used. There are also unsupervised approaches, such as clustering-based methods, that do not require annotations.

An unsupervised learning technique called clustering is used to group related objects according to the maximisation of within-class similarity and the minimization of inter-class similarity. The Manhattan distance or Euclidian distance, respectively, is used as the foundation for the similarity function.

The Euclidian distance can be measured using the formula,

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+\ldots+|x_{i_p}-x_{j_p}|^2)}$$

The operation of K-Medoids, a method for unsupervised clustering based on partitions where each cluster is represented by a single pixel, is explained in the following.

• Create k-nonempty subsets from the pixels.
• Calculate the cluster's seed point. The pixel in the centre of that cluster is identified as Seeds/Medoids.
• Assign every pixel to the cluster with the closest seed.
• Keep going until there are no more assignments.
• Each cluster designates a certain component of the face image.

1.

d) Image Enhancement:

Image enhancement is a process of enhancing an image's visual quality through the adjustment of its contrast, brightness, sharpness, colour balance, and other characteristics. Before images are used for computer vision tasks like object detection, classification, and segmentation in AI, they must first go through an important pre-processing stage called image enhancement.

There are several techniques used in AI for image enhancement, including:

Histogram equalization: This technique adjusts the brightness and contrast of an image by transforming the histogram of the image so that it has a uniform distribution of intensities. This results in an increase in the overall contrast of the image.

Sharpening: By widening the gap between consecutive pixels, this approach improves the borders and fine details of an image. It is helpful for giving the appearance of greater clarity and detail in an image.

De-noising: This technique removes noise from an image by filtering out small variations in intensity values. It is useful for removing unwanted artefacts from an image and improving its clarity.

Colour correction: This technique adjusts the colour balance of an image by changing the levels of red, green, and blue in the image. It is useful for correcting the colour cast in an image and making it more visually appealing.

IV. Classification by CNN:

Convolutional Neural Networks (CNNs) can be implemented in Machine Learning (ML) for the detection of breast cancer. CNNs are well-suited for image classification tasks, and have been used in medical image analysis to classify mammogram images as benign or malignant.
Each layer will have multiple units. The input values to the neural network are the attributes of every training tuple.

Data collection and pre-processing: This involves collecting a large dataset of images that are representative of the different classes to be classified and pre-processing the data by resizing the images to a standard size and converting them to a numerical representation.

Data augmentation: Techniques for data augmentation including random cropping, flipping, and rotation can be used on the photos to expand the dataset size and avoid overfitting.

Designing the CNN architecture: This involves selecting the number of layers, type of layers (such as convolutional, activation, and pooling layers), and the size of filters and pooling regions, based on the size and complexity of the dataset and the desired accuracy.

Training the network: After then, the network is trained on the pre-processed and enhanced data using an optimisation approach to minimise the loss function, such as Stochastic Gradient Descent (SGD) or Adam.

Testing the network: After training, the network is tested on a separate test dataset to evaluate its performance.

Deployment: If the network performs well on the test data, it can then be deployed for use in a real-world setting.

It is crucial to remember that the architecture and specifics of the training process will vary depending on the problem at hand and the data that are available, and that this may necessitate numerous iterations and fine-tuning to attain the best results. The capacity of CNNs to automatically learn hierarchical representations of the input data has enabled them to achieve cutting-edge performance in a range of applications, including object recognition and scene classification.

Figure 3: Convolutional Neural Network

In machine learning and deep learning, the optimization algorithm stochastic gradient descent (SGD) is used to decrease a loss function. By measuring the discrepancy between the expected output and the actual target as measured by the loss function, SGD is used to determine the set of parameters (weights and biases) that result in the lowest loss. SGD works by gradually altering the parameters in the direction of the parameters' obverse gradient of the loss function. The loss function's derivative with respect to the parameters, or gradient, demonstrates how the loss changes as the parameters are altered.

Algorithm:

A loss function L() is minimised with regard to the parameters using the Stochastic Gradient Descent (SGD) algorithm. It is an iterative process, and each iteration updates the parameters using the following formula:

$$\theta = \theta - \alpha * \nabla L(\theta)$$

Here, α is the learning rate and $\nabla L(\theta)$ is the gradient of the loss function with respect to the parameters θ.

An unrelated training sample (x, y) is randomly selected from the training dataset for each iteration, and the gradient of the loss with respect to the parameters is computed using the method below:

$$\nabla L(\theta) = (1/m) * \sum (y\hat{} - y) * x$$

where m is the number of training examples, y^ is the predicted output, y is the target output, and x is the feature vector.
In the direction of a negative gradient, which would reduce the value of the loss function, the gradient is utilised to update the parameters. Until a halting requirement is satisfied or the loss function achieves a minimum, the process is repeated.

IV.CONCLUSION:

Recent researches reveal understanding the importance of machine learning in medical field to detect many diseases. This paper gives an approach of using Machine learning algorithm in detecting Breast Cancer in earlier stage. The idea presented in this paper predicts how successful and ahead we can be in the medical field and save many lives by early detection of Cancer. Machine Learning has shown great promise in the early detection of breast cancer by analysing medical images such as mammograms.

Supervised learning methods, such as deep convolutional neural networks (CNNs), have been widely used for breast cancer detection, and have demonstrated high accuracy in detecting breast tumours. Additionally, unsupervised learning methods, such as clustering algorithms, are also used to identify suspicious regions in mammograms that require further examination by radiologists.

V.REFERENCES

[1] https://www.sciencedirect.com/science/article/pii/S1877050921014629

[2] https://www.researchgate.net/publication/341508593_BREAST_CANCER_PREDICTION_USING_MACHINE_LEARNING

[3] https://www.researchgate.net/publication/347094703_Breast_Cancer_Prediction_Using_Machine_Learning

[4] https://www.hindawi.com/journals/jhe/2022/4365855/

[5] https://www.ijraset.com/research-paper/breast-cancer-prediction-using-machine-learning

[6] https://ieeexplore.ieee.org/document/8862533

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/

[8] https://www.frontiersin.org/articles/10.3389/fpubh.2022.924432/full

[9] https://www.ijert.org/research/breast-cancer-detection-using-machine-learning-techniques-IJERTV10IS070064.pdf

[10] https://www.nature.com/articles/s41598-019-48995-4

[11] https://medium.datadriveninvestor.com/breast-cancer-detection-using-machine-learning-475d3b63e18e

[12] https://arxiv.org/ftp/arxiv/papers/2203/2203.04308.pdf

[13] https://www.hindawi.com/journals/cin/2022/6333573/

[14] https://www.hindawi.com/journals/jhe/2019/4253641/

[15] https://ieeexplore.ieee.org/abstract/document/8769187

[16] Sri Hari Nallamala, Siva Kumar Pathuri, Dr Suvarna Vani Koneru, “A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment”, International Journal of Engineering & Technology (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7 (2018), SI 7, P. 729 – 732.

[17] https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10

[18] https://jtec.utem.edu.my/jtec/article/view/4706

VI. AUTHORS PROFILE

Vibha S has received her BCA degree from Bangalore University in 2020 and is pursuing Master of Computer Application in BMS college of engineering. She has published paper in cloud computing and IoT and it has been published online in IJAECE. She has attended seminars on networking, machine learning and cloud computing. Her areas of interest are Artificial Intelligence, Cloud Computing and Machine Learning.

Shailaja K.P received her B.Sc degree from Bangalore University in 1995, MCA degree from Bangalore University in 1999 and M.Phil degree from BharatiDasan University in 2006 and currently working as Assistant Professor in Department of Computer Applications, B.M.S. College of Engineering, Bangalore. At present she is pursuing her PhD under the guidance of Dr. Manjunath M in the Department of Master of Computer Applications, RVCE, Bangalore from VTU, Belgaum. Her areas of interest are Data Mining and Analytics and Machine Learning.

Pushpa T S received her B.Sc degree from Bangalore University in 1992, MCA degree from Bangalore University in 1998 and M.Phil degree from Madurai Kamaraj University in 2008 and currently working as Assistant Professor in Department of Computer Applications, B.M.S. College of Engineering, Bangalore. At present she is pursuing her PhD under the guidance of Dr. K Vijaya Kumar, Department of Master of Computer Applications, B.M.S. College of Engineering, Bangalore from VTU, Belgaum. Her areas of interest are Data Mining, Data Analytics and Machine Learning.