Breast Cancer Prediction using Machine learning and Deep learning algorithms

M Visleshana
Department of Computer Science and Engineering
B V Raju Institute of Technology
Narsapur, Medak District, Telangana State 502313, India.

L Pallavi
Department of Computer Science and Engineeing
B V Raju Institute of Technology
Narsapur, Medak District, Telangana State 502313, India.

M Basavaprasad
Department of Computer Science and Engineering
B V Raju Institute of Technology
Narsapur, Medak District, Telangana State 502313, India.

Santhosh
Department of Computer Science and Engineering
B V Raju Institute of Technology
Narsapur, Medak District, Telangana State 502313, India.

***Abstract*—One of the diseases with a rapid spread that kills women in their younger years is breast cancer. Machine learning methods can significantly aid in the early detection and prediction of breast cancer. Unfortunately, the patient's life expectancy is reduced because cancer is discovered at a later stage. Their lifespan may have been extended if the detection was made earlier. Therefore, the project attempts to use Macine learning algorithms to predict the presence of breast cancer at an early stage. Machine learning algorithms like Logistic Regression(LR), Support Vector Machine(SVM)(linear), Random Forest(RF), and Naive Bayes(NB) to predict whether the tumor is malignant or benign and also predict breast cancer using Deep learning algorithms like CNN and ANN**

***Index Terms*—Malignant, Benign, Logistic Regression, Support Vector Machine, Random Forest, Naive Bayes, ANN, CNN, Machine learning, Deep Learning.**

I. INTRODUCTION

Breast cancer is a type of cancer that originates in the cells of the breast. It is one of the most common cancers among women worldwide, although it can also occur in men, although it is relatively rare. Breast cancer typically starts in the cells that line the milk ducts (ductal carcinoma) or the lobules (lobular carcinoma) of the breast.

Breast cancer can range from early stage, when it is confined to the breast tissue, to advanced stage, when it has spread to nearby lymph nodes or other distant organs. It can also be classified as invasive, where it has penetrated into the surrounding breast tissue, or non-invasive, where it remains confined to the ducts or lobules without invading the surrounding tissue.

There are several risk factors associated with breast cancer, including age, gender, family history, genetic mutations (such as BRCA1 and BRCA2), hormonal factors, lifestyle factors (such as obesity, lack of physical activity, and alcohol consumption), and exposure to estrogen over a prolonged period (such as early onset of menstruation, late menopause, or hormone replacement therapy).

Common symptoms of breast cancer include a lump or mass in the breast or underarm area, changes in breast size or shape, nipple changes (such as inversion, discharge, or

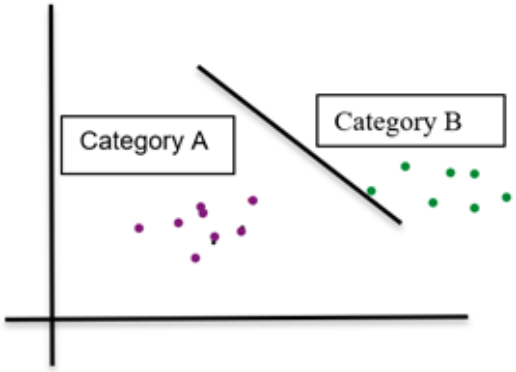

Fig. 1.

redness), breast pain or tenderness, skin changes (such as dimpling, puckering, or rash), and swelling or lumps in the lymph nodes under the arm.

Early detection through regular breast self-examination, mammography, and clinical breast examination is crucial for increasing the chances of successful treatment. Treatment options for breast cancer may include surgery (such as lumpectomy, mastectomy, or lymph node removal), radiation therapy, chemotherapy, hormonal therapy, targeted therapy, or a combination of these approaches, depending on the stage, type, and characteristics of the breast cancer.

Breast cancer is a complex disease that affects millions of people worldwide, and it requires a comprehensive approach to diagnosis, treatment, and support. It is important to consult with qualified healthcare professionals for accurate diagnosis, personalized treatment plans, and emotional support throughout the breast cancer journey. Breast cancer is a terrible illness that will affect women everywhere and causes both physical and psychological harm. Quite a few women are also impacted by it. Around 40 thousand women came close to passing away from the disease, and 2 lakh new cases of

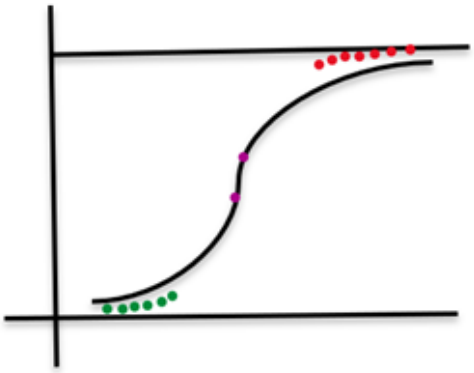

Fig. 2.

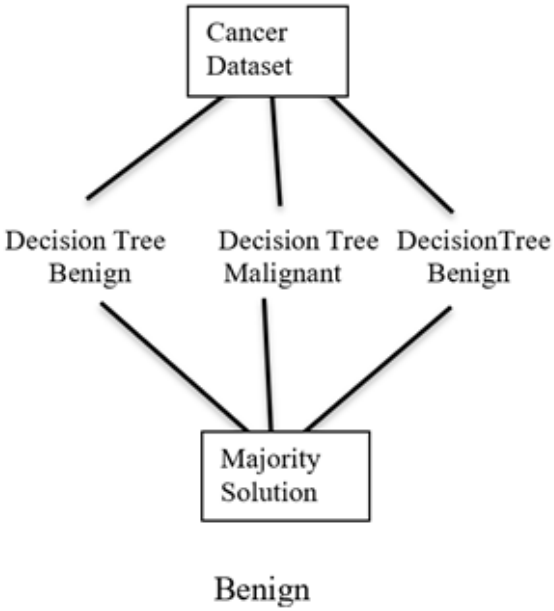

Fig. 3.

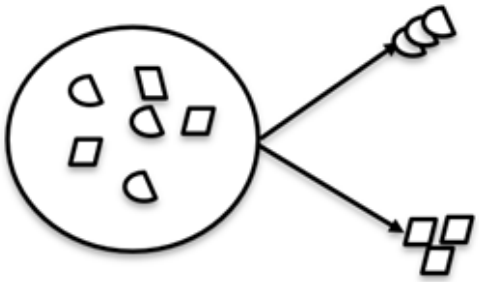

Fig. 4.

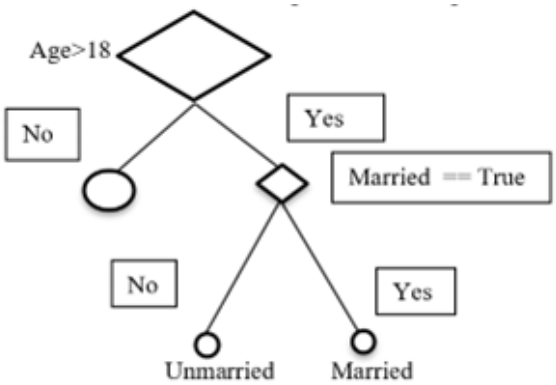

Fig. 5.

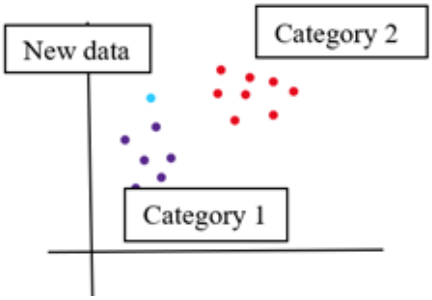

Fig. 6.

breast cancer in women were anticipated. There are various types of breast cancer: one is benign and another one is malignant. Different treatment initiatives for therapy will be made by professionals and doctors depending on the diagnosis of breast cancer. Treatments will lead to ineffective therapies and allow patients to miss the optimum window of opportunity for recovery, both of which will have terrible consequences. Therefore, it is crucial to use a model that can accurately predict the kind of breast tumor.

Nowadays, one in five individuals globally will have cancer at some point in their lives. According to projections, more people will be given cancer diagnoses in the upcoming years,reaching a level that is about 50% higher in 2040 than it was in 2020. As of 2020, there will be 10 million cancer-related deaths worldwide, up from 6.2 million in 2000.Cancer is the cause of more than one in six deaths. This article primarily compares the effectiveness of several classifiers, including Support Vector Machines (SVM), Random Forest(RF), Logistic Regression(LR), Decision Trees(DT), and K-Nearest Neighbors (KNN)

II. LITERATURE SURVEY

[1]Mostafa Shanbehzadeh et al proposed a method to predict breast cancer using machine learning algorithms like Random Forest(RF), Logistic Regression(LR), Naïve Bayes(NB), Decision Tree(DT), Xgboost and Support Vector Machine(SVM)

algorithms by considering parameters like Body Mass Index, Age, Hormone Therapy i.e, Estrogen and progesterone, Breast Feeding, Alcohol/smoking, Hypertension, Hypercholesterolemia, Hyperglycerdemia, Hyperlipidemia, Diabetes, obesity. They stated Random Forest is the best algorithm with 78.869% accuracy among all algorithms. [2]Muhammet Faith Ak et al concluded a model to predict breast cancer using Machine Learning algorithms such as Logistic regression, K-Nearest Neighbour(KNN), support vector machine(SVM), naïve Bayes (NB), and Decision trees(DT) while taking into account the characteristics of the tumor mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, se radius, se texture, se perimeter, se area, se smoothness, se compactness, se concavity, se concave points, se symmetry, se fractal dimension, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points, worst symmetry, worst fractal dimension. They stated Logistic Regression is the best algorithm with 98.1% accuracy. [3] Noreen Fatima et al inferred a method to Predict the breast cancer using Machine Learning algorithms like Artificial Neural Network (ANN), Logistic Regression(LR), K-Nearest Neighbour(KNN), Decision Tree(DT), Naïve Bayes(NB), Support Vector Machine(SVM), Random Classification And Regression Tree by considering the data from pathology report and WDBC. They stated Support Vector Machine Algorithm is the best algorithm with 98.03% accuracy. [4]Tuan Tran et al suggested a model for Breast Cancer Prediction using Machine Learning Algorithms like K-Nearest Neighbour(KNN), Neural Network, Decision Tree(DT), Random Forest(RF), Support Vector Machine(SVM) by considering attributes like Clump Thickness, Cell size, Cell shape, Adhension, Nuclei, Chromatin, Mitoses. They Stated XGB Tree achieved the best performance with 97.7% accuracy.

[5]Yixuan Li et al proposed a model for Breast Cancer Prediction using Machine Learning Algorithms like Decision Tree(DT), Support Vector Machine(SVM), Random Forest(RF), Logistic Regression(LR), and Neural Network by considering the data from BCCD data set. They stated Random forest is the best algorithm with 93.7% accuracy.

TABLE I
ALGORITHM ACCURACIES

S.NO	Authors	Best Algorithm	Accuracy
1	Mostafa Shanbehzadeh	Random Forest	78.869%
2	Muhammet Faith Ak	Logistic Regression	98.1%
3	Noreen Fatima	Support Vector Machine	98.03%
4	Tuan Tran	XGB Tree	97.7%
5	Yixuan Li	Random Forest	93.7%

III. EXISTING SYSTEM

Data Collection: Relevant data, such as patient demographics (e.g., age, sex), clinical history (e.g., family history of breast cancer), and imaging or pathological data (e.g., mammogram images, biopsy results), are collected from various sources, such as electronic health records (EHRs), medical databases, or research studies. Data Preprocessing: The collected data may need to be preprocessed to ensure data quality and consistency. This may involve cleaning and handling missing data, normalizing or standardizing feature values, and encoding categorical variables into numerical representations. Feature Selection: Relevant features, or variables, that are most informative for breast cancer prediction are selected from the preprocessed data. This may involve statistical analysis, domain knowledge, or feature importance ranking using techniques like feature selection algorithms or dimensionality reduction techniques. Model Training and Validation: Machine learning models, such as logistic regression, decision trees, random forests, support vector machines, or deep learning algorithms, are trained on a portion of the preprocessed data. The remaining data is used for model validation to assess its performance. Techniques like cross-validation may be used to evaluate the model's performance on different subsets of the data. Model Evaluation: The trained model is evaluated using appropriate evaluation metrics, such as accuracy, sensitivity, specificity, precision, recall, F1 score, or area under the receiver operating characteristic (ROC) curve. This helps assess the model's predictive accuracy and generalizability to new, unseen data. Model Deployment: Once the model is deemed

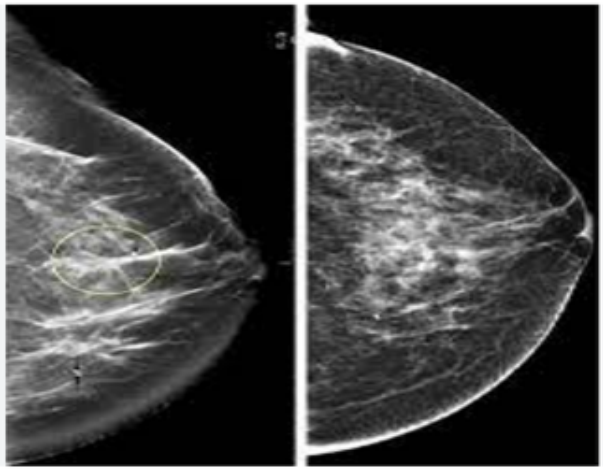

Fig. 7.

accurate and reliable, it can be deployed in a real-world clinical setting. This may involve integrating the model into an electronic health record (EHR) system, a decision support tool, or a mobile application for clinical use. A. Drawbacks • The existing system cannot predict cancer by directly using images as a dataset. • The suggested system can compare all the algorithm's accuracy and picks the best accuracy. • The suggested system can predict through images by using CNN and ANN algorithms.

IV. PROPOSED SYSTEM

This paper proposes a continuous traffic information collection system based on video-based vehicle identification, classification, counting, and speed estimation. The system

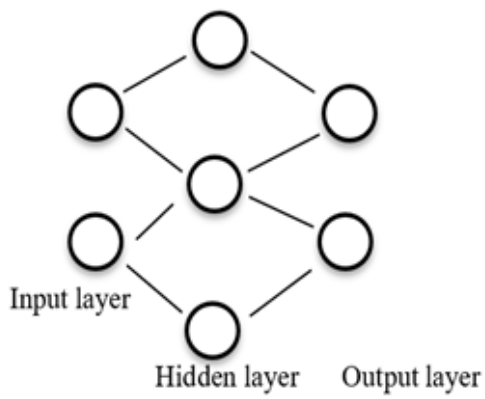

Fig. 8.

was developed using Python (https://python.org/) and OpenCV (https://opencv.org). The main objective of this system is to gather data on vehicle counts and characteristics, which can be used to create an intelligent transportation network based on historical traffic data. By detecting, classifying, counting, and estimating the speeds of vehicles, the proposed system can generate traffic information, which is saved in CSV/XML file format. The MOG2 algorithm is used for background subtraction in this plug-and-play system. The proposed system was tested in six locations across Delhi, India's capital, under various traffic and environmental conditions. It accurately classified all types of vehicles 81% of the time in Delhi. The rest of this paper is organized as follows: Section II describes the approach used to build the proposed system, Section III describes the proposed system's graphical user interface (GUI), Section IV presents the results of the experimental tests, and Section V concludes with conclusions and suggestions for future research.

V. METHODOLOGY

1. Through Excel data using 32 attributes 2. Through images using Deep Learning algorithms – CNN, ANN

TABLE II
MEAN FEATURES

ID	Mean radius	Mean texture	Mean parameter
1	Mean area	Mean smoothness	Compactness
2	Mean symmetry	Concave mean points	Mean fractal dimension
3	Radius se	se Texture	se Perimeter
4	se area	Compactness se	Smoothness se
5	Compactness se	se Concavity	se Concave points

Methodology – I: The breast cancer dataset contains 32 parameters id Mean radius Mean texture Mean parameter Mean area mean smoothness Compactness mean mean symmetry Concave mean points Mean fractal dimension Radius se se Texture se Perimeter se area Compactness se Smoothness se Compactness se se Concavity se Concave points These 32 parameters are used to classify the tissues either benign or malignant. If the parameter values are relatively higher then it indicates that contains malignant tissue i.e, cancerous tissue. In this methodology, we build different models using different machine learning algorithms like SVM, Logistic Regression, Random Forest, Naïve Bias, KNN, XGBOOST, CART Support Vector Machine: A supervised learning environment algorithm. In this algorithm best boundary line is created i.e, Hyperplane, this line separates the space into categories. Such that, a given data point is inserted in the correct position.

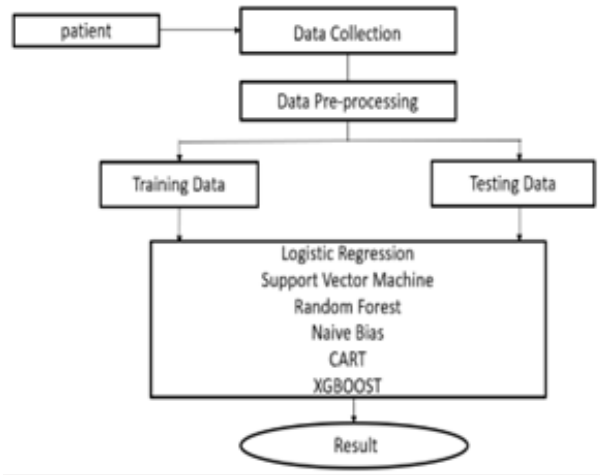

Fig. 9.

Logistic Regression: It is used to solve classification problems, it predicts the output i.e, dependent variable based on independent variables. The predicted output is a categorical value.

Random Forest: It is used to solve classification and Regression problems. This algorithm uses a classifier to generate multiple decision trees on the dataset to predict the output.

Naive Bias: It is employed to address classification issues. In this algorithm, a probabilistic classifier is used to predict the output based on the probability of features.

Decision Tree: It is utilised for solving Regression and Classification problems. The nodes of the tree represent the columns of dataset, braches represent the rules of decision tree and the leaf node of the decision tree represents the output.

K- Nearest Neighbour: A supervised learning environment algorithm. In this algorithm, the new data point observes the relation between already trained data then, new data is inserted into similar category.

XGBOOST: A supervised learning environment algorithm. It is utilised for solving classification and regression problems. It uses to build shallow decision trees sequentially to predict accurate results.

Classification And Regression Tree: It is used to solve classification problems. In this algorithm, the data set features are splitted into classes. The decision tree is split into predictor variables, which have the prediction of the target variable.

Methodology – II

In this methodology, the data set contains the images i.e, Mammograms which means x-ray images of breast. Breast

cancer is a serious health issue that affects millions of women worldwide. Medical imaging plays a crucial role in the early detection and diagnosis of breast cancer. However, the accuracy of computer-aided diagnosis systems, such as CNNs, heavily depends on the quality of the images used for training and testing.

The above image is mammographic image i.e, x-ray and the set of images to be trained under CNN and ANN algorithms.

To improve the performance of CNNs, various image processing techniques can be applied to breast cancer images. They are:

Image resizing: It can help to standardize the size of the images and reduce computational complexity.

Image normalization: It can help to standardize the pixel values and reduce the effects of illumination and contrast differences.

Image augmentation: It can be used to generate new training samples and improve the generalization ability of CNNs. Image denoising can help to remove noise from the images and improve the accuracy of CNNs.

Image enhancement: This technique can be used to improve the visual quality of breast cancer images, making it easier to identify suspicious regions.

Image segmentation: This technique can be used to isolate the breast region and suspicious regions for further analysis.

By applying these image processing techniques, the accuracy and performance of CNNs and ANN can be significantly improved, leading to more accurate and reliable breast cancer diagnosis and treatment.

CNN: Convolutional neural networks (CNNs) are a particular family of artificial neural networks that are particularly effective at classifying images and identifying objects. The fundamental principle of CNNs is the use of convolutional layers, which are layers of learnable filters that may extract ever more complicated characteristics from the input image.

The steps in the CNN algorithm typically proceed as follows:

Convolutional Layers: In the network's first layer, a set of filters that each extract a different feature from the input image are convolved. A collection of feature maps that depict the presence of various features in the image are the layer's output.

Pooling Layers: Usually, the convolutional layer's output is followed by a pooling layer, which lowers the feature maps' dimensionality while keeping the most crucial characteristics. Max pooling, which takes the maximum value inside a region of the feature map, is the most used kind of pooling procedure.

Fully Connected Layers: To create the network's final output, the output of the pooling layer is flattened and sent through one or more fully connected layers. These layers then conduct a linear transformation on the features.

Activation function: An activation function, such as the Rectified Linear Unit (ReLU) function, which brings nonlinearity into the model, is often included in each layer of the CNN.

Loss Function and Optimization: The CNN is trained using a loss function that assesses the discrepancy between the true label for a specific input image and the predicted label. The network's weights are subsequently updated using an optimisation approach, such as stochastic gradient descent, in order to reduce the loss function.

Testing: After being trained, the CNN may be used to categorise fresh input images by running them through the network and choosing the category with the highest output probability.

ANN: The structure and operation of the human brain served as the basis for the development of Artificial Neural Networks (ANNs), a type of machine learning algorithm. Classification, regression, and pattern recognition are just a few of the many tasks that ANNs excel at.

An ANN's input layer, one or more hidden layers, and output layer make up its fundamental structure. A group of neurons that are connected to one another and exchange information through weighted connections make up each layer.

The following steps make up the ANN algorithm typically:

Initialization: Random initialization is used to determine the weights of the connections between the neurons.

Forward Propagation: The input data is supplied into the input layer, and each layer's weighted summations and activation functions are used to calculate the neurons' activations. The network's anticipated output is represented by the output of the final layer.

Loss Function and Optimization: A loss function is calculated to calculate the difference between the predicted and true outputs after comparing the network's predicted output with the actual output. The weights of the network are then updated using an optimization approach, such as stochastic gradient descent, to reduce the loss function.

Backward propagation: Through the use of backpropagation, the gradients of the loss function with respect to the network weights are calculated. To minimize the loss function, these gradients are then utilized to update the network's weights in the gradient's opposite direction.

Testing: After the artificial neural network (ANN) has been trained, it can be used to predict the output of new input data by introducing it to the input layer and propagating it through the network.

Steps for predicting Breast Cancer:

Data Gathering: Compile a sizable database of photos from mammograms showing both benign and malignant tumours. The collection should be well-balanced and include a wide range of photos that represent various stages and types of breast cancer.

Data preprocessing: To enhance the quality of the photographs, preprocess the images by resizing, standardizing, and enhancing the data.

Model Architecture: Create a model architecture for a feedforward neural network that analyses input photos and forecasts the likelihood of breast cancer. There should be an input layer, a few hidden layers, and an output layer in the model. Equal to the number of features in the input data, the input layer should have the same number of nodes. (i.e., the number of pixels in the image). A single node that

outputs from the output layer is required. Or Create a model architecture for Convolutional neural networks.

Model Training: Apply the preprocessed data to the ANN model training. Improve the performance of the model by using methods like regularization, gradient descent, and back-propagation.

Model Evaluation: On the test dataset, evaluate the model's performance using measures like accuracy, precision, recall, and F1-score.

In addition to the features mentioned above, CNNs can also learn to recognize other characteristics associated with breast cancer, such as the presence of lymph nodes or the shape and size of the tumor. CNN accomplishes this recognition by analyzing the images using convolution to identify areas of the image that match a particular pattern or feature. The CNN then combines these features and analyzes them in the fully connected layers of the network to make a final prediction about the presence of cancer in the image.

To train the CNN, a large dataset of labeled breast images is required. This dataset is used to train the CNN to recognize the patterns and features associated with cancer. The CNN is then tested on a separate dataset to evaluate its performance and accuracy.

Overall, using a CNN to analyze breast images for signs of cancer can be an effective tool for early detection and diagnosis. It allows for a more accurate and efficient screening process, potentially leading to earlier detection and better patient outcomes.

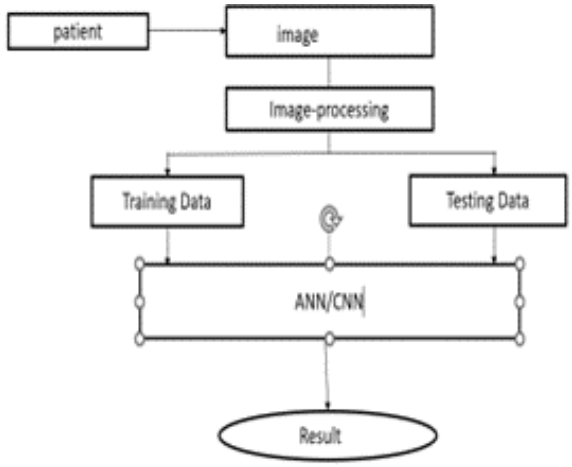

Fig. 10.

For Methodology-I:

In this methodology, the data is collected in the form of excel sheet and 32 parameters are stored in this data set. By preprocessing techniques the data will be preprocessed and it splits into training and testing data then, The model will be trained with mentioned any one of the above algorithms and after comparing accuracies the best algorithm is considered among them.

For Methodology-II

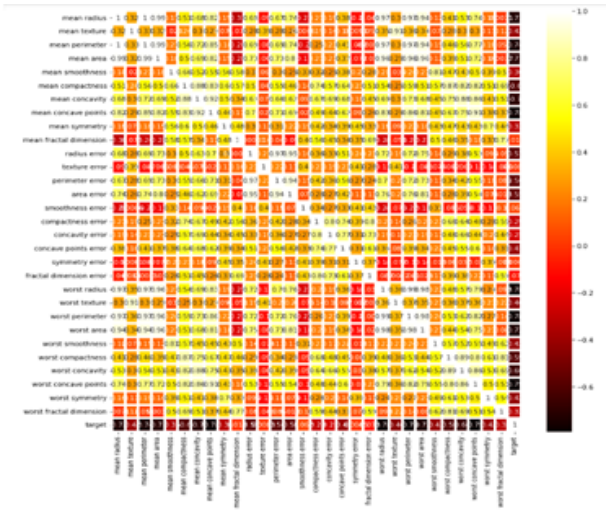

Fig. 11.

In this methodology, the data is collected in the form mammographic x-ray image, by image preprocessing techniques the images are preprocessed then. The model is trained with ANN/CNN then the accuracies will be compared.

VI. RESULT AND ANALYSIS

For Methodology – I A heatmap was created to see how closely all of the dataset' attributes correlated with one another.

If the value of the target is 1 then it is malignant tissue, if it is 0 then it is benign. We collected datasets from three different hospitals Chelmeda Ananda Rao hospital Karimnagar, MNJ cancer hospital Hyderabad and Cancer Clinics Hyderabad. Algorithms for identified data are part of the supervised learning technique by understanding the given data and generating future predictions. Under this strategy, there are two distinct categories: classification and prediction.Machine learning applications typically favor various classification algorithms. The algorithms accuracies observed in this project are Support Vector Machine(SVM), K-Nearest Neighbour(KNN), Logistic Regression, Random Forest, Naïve bias, Decision Tree, xgboost, Classification, And Regression Tree(CART). The Collected data sets from three different hospitals are split into training data at 75% and testing data at 25%. The logistic regression algorithm produced accuracy results of 98.07%, 97.51%, and 96.65% with the help of man libraries. With the use of the KNN algorithm, accuracy scores were obtained as 92.39%,91.37%, and 90.79% for each dataset.The SVM algorithm was used to obtain accurate results.SVM accuracy was 92.37%, 90.79%, and 89.79%. In comparison to the other methods, the Nave Bayes method produced the worst results. Accuracy for each dataset was 92.63%, 91.88%, and 90.7%.

In ANN accuracy scores observed were 92.36%, 91.29%, and 90.79%. The accuracies observed in CART are 92.79%, 90.25%, and 88.79%. In random forests, the data is built into N decision trees. There are N decision trees in the random forest. In this operation, N was set to 30, and the accuracy

results for each dataset were 93.69%, 92.77%, and 93.97%. The accuracies observed in the decision tree are 90.79%, 91.25%, and 92.79%. The accuracies observed in XGBOOST are 91.79%, 90.25%, and 89.79%. For Methodology – II In methodology – I to determine whether the breast has cancer or not we have to take more parameters. In methodology – II, we just take the Mammograms i.e, the x-ray images of the breast. In this methodology, the images are given as input for CNN and ANN then the accuracies are observed i.e, 93% and 85%.

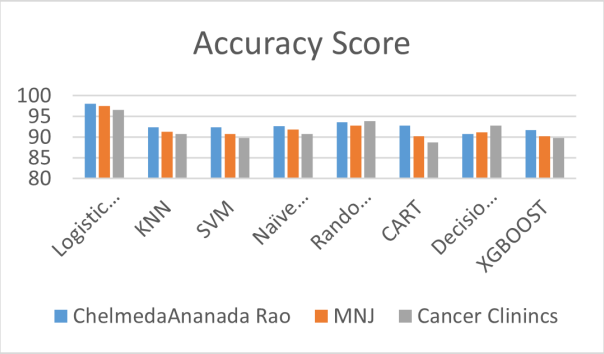

Fig. 12.

VII. CONCLUSION

For Methodology - I The ability to quickly identify breast cancers using various methods that enhances the standard of care and offers patient-specific treatments. The most effective machine learning techniques The RandomForest method was presented in the current study as the best model. Because the logistic regression undergoes overfitting i.e, the machine is not much accurate as human being, So the machine's accuracy is 85 to 95

From the above graph, the accuracy score is highest for logistic regression i.e, 98.07%. The logistic regression undergoes overfitting i.e, the machine is not much accurate as a human being, So the machine's accuracy is 85 to 95 % range which is taken as the best algorithm. So, here next highest accuracy Random forest algorithm is taken as a best algorithm.

For Methodology – II: CNN have highest accuracy than ANN . So for the image processing in breast cancer CNN have highest accuracy than ANN.

VIII. FUTURE WORK

To increase the efficiency of classification algorithms and allow them to make predictions on a wider range of criteria, it is necessary to perform an additional study in this field.In order to attain high accuracy.We are researching a variety of datasets and the potential applications of machine learning techniques to further identify breast cancer. We focus on maximizing accuracy while lowering error rates.

IX. REFERENCES

[1] Mostafa Shanbehzadeh,Hadi Kazemi-Arpanahi ,Mohammad Bolbolian Ghalibaf , Azam Orooji ," Performance evaluation of machine learning for breast cancer diagnosis: A case study ", published: 27 jun 2022, DOI: https://doi.org/10.1016/j.imu.2022.101009 .

[2] Joyce Ayoola; Tokunbo Ogunfunmi, "A Comparative Analysis of Regression Algorithms with Genetic Algorithm In The Prediction of Breast Cancer Tumors", published: sep 2022 , DOI: 10.1109/GHTC55712.2022.9911033.

[3] Krishna Mridha, "Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques", published: jun 2021, DOI: 10.1109/CSNT51715.2021.9509658.

[4] Mohammed Amine Naji , Sanaa El Filali Kawtar Aarika , EL Habib Benlahmar , Rachida Ait Abdelouhahid, Olivier Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis" ScienceDirect , published: aug 2021

[5] Ramik Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING", published: 20 may 2020.

thank our colleagues and collaborators for their support an