International Journal of Electronics and Communication Engineering.
ISSN 0974-2166 Volume 16, Number 1 (2023), pp. 17-30

https://dx.doi.org/10.37624/IJECE/16.1.2023.17-30

Breast Cancer Detection Using Support Vector Machine Algorithm

[1]V. Uma, [2]Varshini Boorla, [3]Likhitha Bonala, [4]Mohammed Sana and [5]Vennamaneni Shrestha

[1]Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student
G. Narayanamma Institute of Technology and Science (For Women)
Hyderabad, India
uma.volety@gmail.com

Abstract

Breast cancer is primary cancer affecting women and ranks second as the leading cause of female mortality. The crucial aspect is identifying the presence of breast cancer and pinpointing the affected area. Medical imaging consistently advances, and early detection of cancer is vital in lowering cancer death rates. The enhancement procedure for mammograms involves filtering and discrete wavelet transforms. Contrast stretching is utilized to boost image contrast. Improved Breast cancer early detection and diagnosis are achieved by segmenting mammogram images. From the segmented breast region, features are retrieved. The proposed system identifies the cancer region and classifies patients as either normal or cancerous. The input mammography image is subjected to pre-processing techniques, and undesirable parts of the image are cropped off. Using morphological techniques, the tumor location is separated from the surrounding tissue and marked on the original mammography image. If the mammogram image is normal, the patient is deemed normal; otherwise, the patient is diagnosed with cancer. The Decision Tree Algorithm is utilized for categorization in this study for reasons of comparison.

Keywords: preprocessing, Feature extraction, Segmentation, Confusion Metrix.

I. INTRODUCTION

Breast tissue cells can become cancerous and grow into breast cancer. It is the highly prevalent cancer in women around the world.

When cells begin to multiply uncontrollably and form a tumour, breast cancer occurs. These tumors may be malignant (cancerous) or benign (non-cancerous). Malignant tumours have the potential to spread to other parts of the body and represent a threat to life if they are not identified and treated in a timely way. Breast cancer detection is the process of determining whether a person has the disease. Early breast cancer detection is crucial because it enables quick treatment and increases the likelihood of a favorable outcome. A number of screening procedures, such as mammography, ultrasound, MRI, and clinical breast exams, can find breast cancer.

Mammography is a quite popular technique to diagnose breast cancer, which involves capturing X-ray images of breast tissue. While MRI creates detailed images of the breast with a magnetic field and radio waves, ultrasound employs high-frequency sound to create breast tissue images. It is important to keep in mind that not all breast lumps or anomalies are malignant. In reality, benign breast masses predominate. But if a lump or other alteration is found, it should be examined by a medical expert to see if more testing or treatment is required.

In order to analyze the dataset, support vector machine algorithms are used to ascertain the correctness of mammography pictures. The decision tree approach is additionally employed for classification, enabling a comparison of the effectiveness of the two algorithms. The dataset has been split into numerous classes by the support vector machine technique according to particular attributes, giving information based on the accuracy of the mammography image. The decision tree technique, contrasted with, iteratively breaks the data down into smaller groups depending on the most useful qualities, ultimately arriving at a choice at each node of the tree.

Researchers can assess the accuracy of classification findings and determine whether the method is more successful at spotting breast cancer in mammography pictures by using both the SVM and decision tree algorithms. To assess each algorithm's performance, various metrics can be assessed, including sensitivity, specificity, and positive predictive value. The suggested system seeks to address two issues in particular. A Gaussian mixture model is employed to address the first issue, which is locating the breast cancer-affected area. The Support Vector Machine (SVM) algorithm and the Decision Tree algorithm are is employed to address second challenge, which is to categorize patients as normal or malignant.

Objectives of the proposed system are as follows:

1. Acquiring input mammogram images from the dataset.
2. Pre-process the images using the median filter.
3. Extracting features such as mean, variance, and Gabor features.
4. Segmenting the pre-processed mammogram image.
5. Classifying the input images using both the Decision Tree algorithm and the Support Vector Machine (SVM) algorithm.
6. Analyzing performance metrics, such as accuracy and precision.
7. Comparing the output matrix obtained using the Decision Tree algorithm and the Support Vector Machine algorithm.

II. RELATED WORKS

According to a study by Vaishnavi Patil, Shravani Burud, Goutami Pawar, Tanaya Rayajadhav, and Sunil B. Hebbale, [1] undesirable elements and fluctuations in the scanned images, such as noise and brightness variations, can be eliminated. Through the use of an image preparation technique, they have eliminated these undesirable elements.

Vishal Deshwal and Mukta Sharma [2] created a grid search approach for identifying breast cancer in this journal. Without a grid search, the support vector machine model is initially assessed. Grid search is then used to test the support vector machine model. Finally, a comparison study was conducted, and a new model was developed in light of the results. Before fitting it for prediction, the new model is based on a data grid search, which enhances the results.

An algorithm for tumor detection has been proposed by Y. Irenaeus Anna Rejani et al [3].

A breast CAD technique based on feature fusion with convolutional neural network (CNN) deep features was studied by Zhiqiong Wang, Mo Li, Huaxia Wang, Hanyu Jiang, Yudong Yao, Hao Zhang, and Junchang Xin [4]. First, a mass identification technique based on unsupervised extreme learning machine (ELM) clustering and CNN deep features were suggested. Furthermore, a feature set was created by combining deep features, morphological characteristics, texture features, and density features. The ELM classifier is then created to differentiate between benign and malignant breast tumors.

An efficient AdaBoost algorithm for early breast cancer diagnosis and detection was proposed by Jing Zheng, Denan Lin, Zhongjun Gao, Shuang Wang, Mingjie He, and Jipeng Fan [5].

A method with two primary parts has been proposed by M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah [6]. In the first phase, image processing techniques are used to prepare the mammography pictures for feature and pattern extraction. The collected features are used as input for the Back Propagation Neural Network (BPNN) model and the Logistic Regression (LR) model, two different supervised learning models, in the following part.

By analyzing hostile ductal carcinoma tissue zones in whole-slide images, Saad Awadh Alanazi, M. M. Kamruzzaman, Md Nazirul Islam Sarker, Madallah Alruwaili, Yousef Alhwaiti, Nasser Alshammari, and Muhammad Hameed Siddiqi [7] proposed a technique to improve the automatic identification of breast cancer.

III.THE PROPOSED MODEL

The method of detecting breast cancer involves several key steps. First, an input image is collected and pre-processed to prepare it for image processing and segmentation. Next, the image is segmented to isolate the region of interest Features are taken from the image following segmentation, which are then used to train a classifier model. Once the model is trained, It can be utilized to foresee the status of new images. Overall, the breast cancer detection process involves an assortment of imaging methods, feature

extraction, and machine learning algorithms to accurately identify breast cancer and inform treatment decisions.

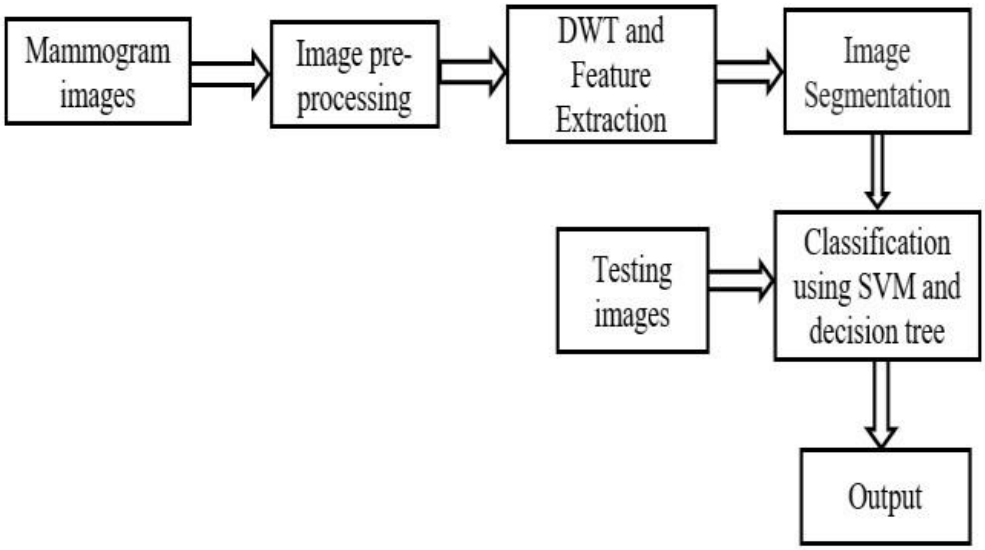

Figure 1: Block Diagram of Proposed system.

3.1 Datasets

The machine learning models utilized in breast cancer detection require mammogram images for training purposes. There are two chosen datasets for use in all the models. Two sets of these datasets are created One has more images for training, while the other has fewer images for testing. These datasets are chosen with care to make sure that they contain a representative sample of mammogram images and that the models are trained and tested on a diverse range of images. By using these datasets, the models can be trained and tested in a consistent and reproducible manner, allowing for the accurate evaluation of their performance.

TABLE 1 Datasets

S. No	Dataset	Number of classes	Total images	Training images	Testing images
1.	Dataset 1	2 (Benign &Malignant)	1728	1382	346
2.	Dataset 2	3 (Benign, Malignant & Normal)	120	96	24

Table-1 illustrates each class's total number of images, together with the total number of training and testing photographs.
Two datasets were utilized to train and test machine learning models for detecting breast cancer consist of mammograms breast images. Dataset 1 contains a total of 1728 mages, with 2 classes-benign and malignant, with 1382 images used for training (691 images per class) and 346 images for testing (173 images per class). Dataset 2 is a collection of breast images, containing 3 classes with a total of 60 images, where 48 images (16 images per class) are used for training and 12 images (4 images per class) are used for testing. The larger dataset (Dataset 1) is primarily used for training the models, while the smaller dataset (Dataset 2 is employed to assess the performance of the models. These datasets provide a valuable resource for developing machine learning models that can correctly differentiate between benign and cancerous breast pictures.

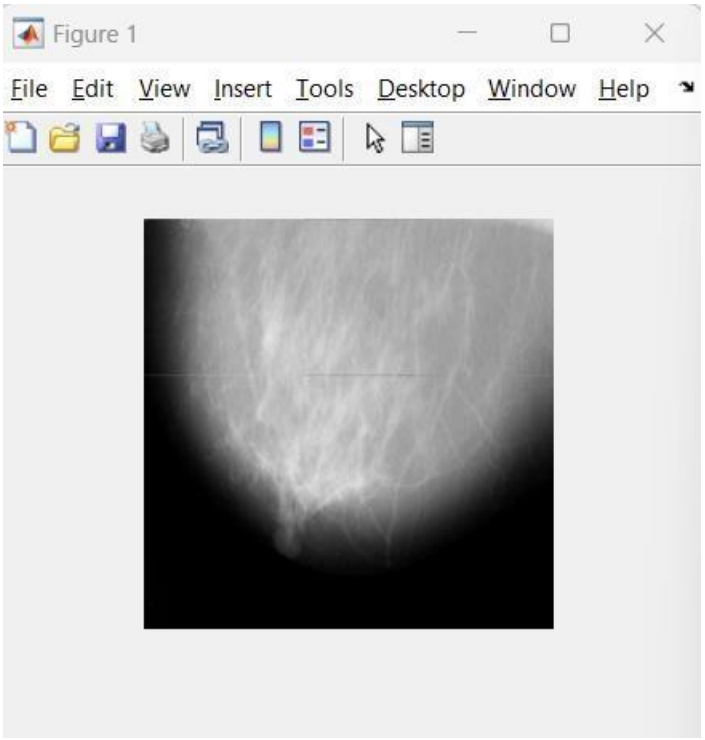

Figure 2. Input Image

3.2 Pre-processing

The first dataset, referred to as dataset 1, has images that have already undergone CLAHE augmentation. As a result, we only need to apply a median filter as part of the image preprocessing. The objective of using a median filter in MATLAB is for the purpose of eliminating any image noise. This filter is non-linear and works by smoothing the image.

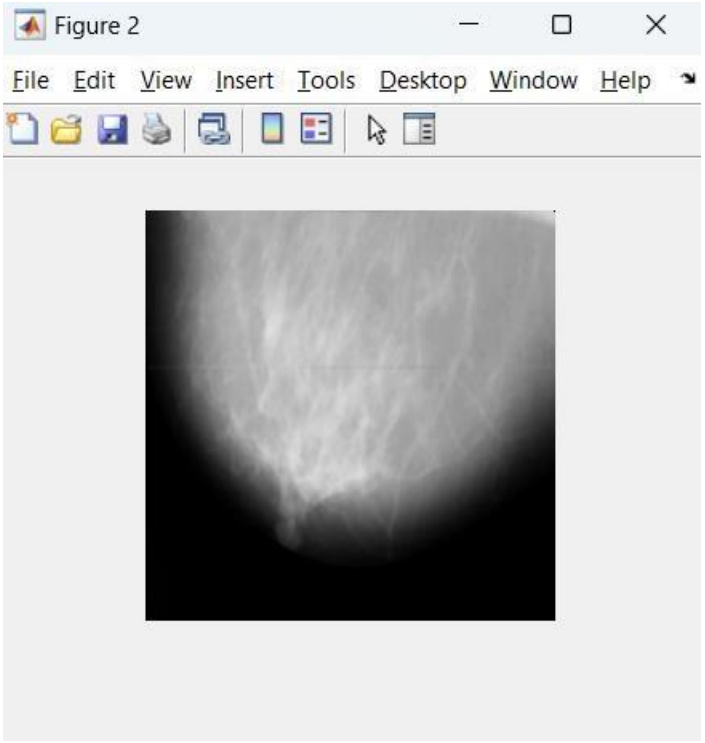

Figure 3. Pre-processing Image
y(t) = median(x(t-T/2), …….x (t), …….x(t+T/2).

3.3 DWT and Feature Extraction

3.3.1 DWT (Discrete Wavelet Transform)

The Discrete Wavelet Transform (DWT) is a mathematical tool used in image processing for analyzing and processing images It breaks down an image into a collection of wavelet coefficients that stand in for various frequency sub bands at various scales. This allows for efficient compression, denoising, and feature extraction of images. The DWT is widely used in applications such as image compression, denoising, and obtaining features for object recognition and classification.

3.3.2 Feature Extraction

Gabor features

In order to analyze and categorize mammography pictures for breast cancer detection, Gabor characteristics have been used. By extracting Gabor features from mammogram images, it is possible to represent the breast tissue in a way that is suitable for machine learning algorithms. Gabor features have been shown to be effective in capturing the texture and frequency characteristics of breast tissue in mammograms. They can be used to classify mammogram images as either benign or malignant, helping radiologists to make more accurate diagnoses cut down on the quantity of false positives and false negatives.

Performance matrices

Although accuracy is a popular performance indicator for classification issues, machine learning models are also frequently evaluated using a number of other metrics. Here are seven commonly used metrics:

1. Accuracy: The proportion of correct predictions out of total predictions made by the model.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

2. Precision: The proportion of true positive predictions out of all positive predictions made by the model.

$$Precision=TP/(TP+FP)$$

3. Recall/Sensitivity. Out of all real positive cases, the percentage of predictions that came true.

$$Sensitivity=TP/(TP+FN)$$

4. F1 score: A combination of precision and recall that considers both metrics.

5. Confusion matrix: A table that lists a categorization model's effectiveness by comparing actual values to predicted values. The number of real positives, fake positives, real negatives, and fake negatives are displayed.

The selection of the metric(s) to employ is dependent on the specific problem being addressed, and the tradeoffs between different metrics must be carefully considered.

3.4 Segmentation

Gaussian mixture models (GMM) have been used in breast cancer detection to segment mammogram images. The segmentation process involves separating the breast tissue from the background and then identifying regions of interest, such as potential tumors. GMM is a clustering algorithm that can model the distribution of an image's pixel intensities. By fitting a GMM to the pixel intensities in a mammogram, it is possible to identify different tissue types and segment the image into distinct regions. This segmentation can aid in the detection and diagnosis of breast cancer by identifying potential tumors and providing more accurate measurements of tumor size and location.

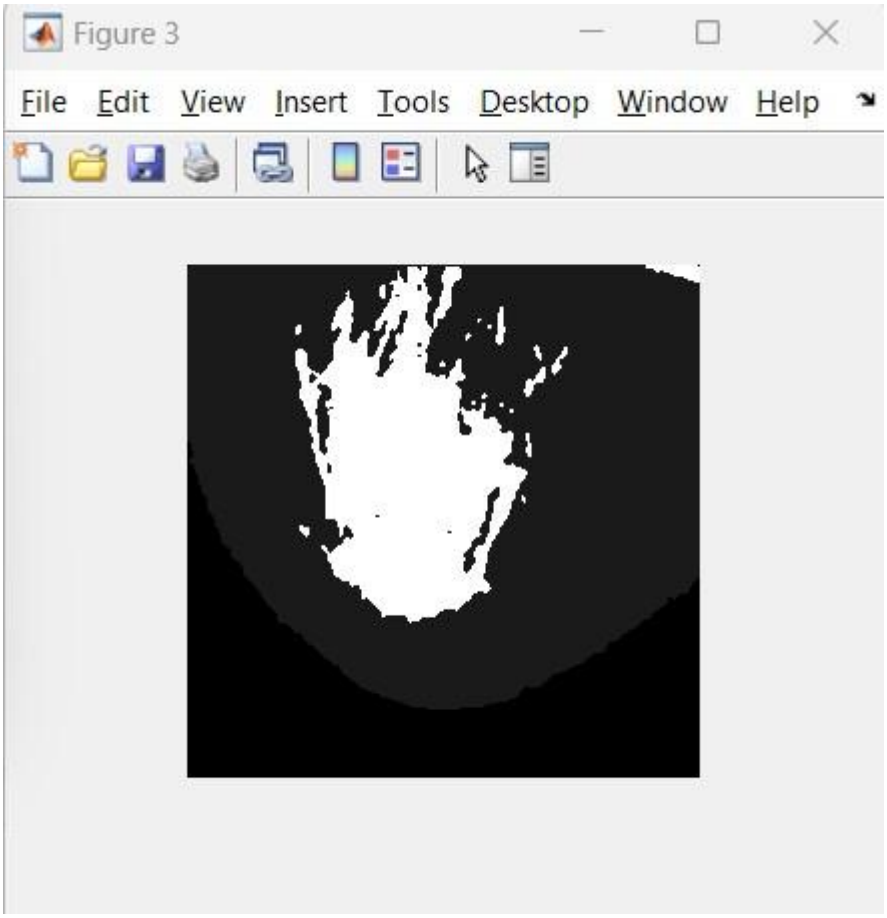

Figure 4. Segmented Image

IV. Classification of Cancer

Support vector machine algorithm and decision tree algorithm are two methods by which the classification of cancer can be accomplished.

4.1 Support Vector Machine Algorithm

The support vector machine (SVM) algorithm is a powerful machine-learning technique that can be used to aid in the detection and classification of breast cancer.

In the context of breast cancer detection, SVMs can be trained on large datasets of breast cancer cases, along with various clinical and demographic variables such as patient age, tumor size, and histological features. The algorithm learns to classify cases as either benign or malignant based on patterns in these variables. Once the SVM is trained, it can be used to analyze new breast cancer cases and make predictions about the likelihood of malignancy. The SVM can also be used to identify important features that contribute to the classification, which can aid in the development of new diagnostic tools and treatments.

4.2 Decision Tree Algorithm

The decision tree algorithm is another machinelearning technique that can be used for breast cancer detection.

The algorithm is trained on a dataset of breast cancer cases and clinical variables and learns to classify cases as either benign or malignant based on patterns in these variables. Decision tree algorithms can also identify important features that contribute to the classification and can be easily interpreted by healthcare professionals, aiding in decision-making, and improving patient outcomes.

V. RESULTS

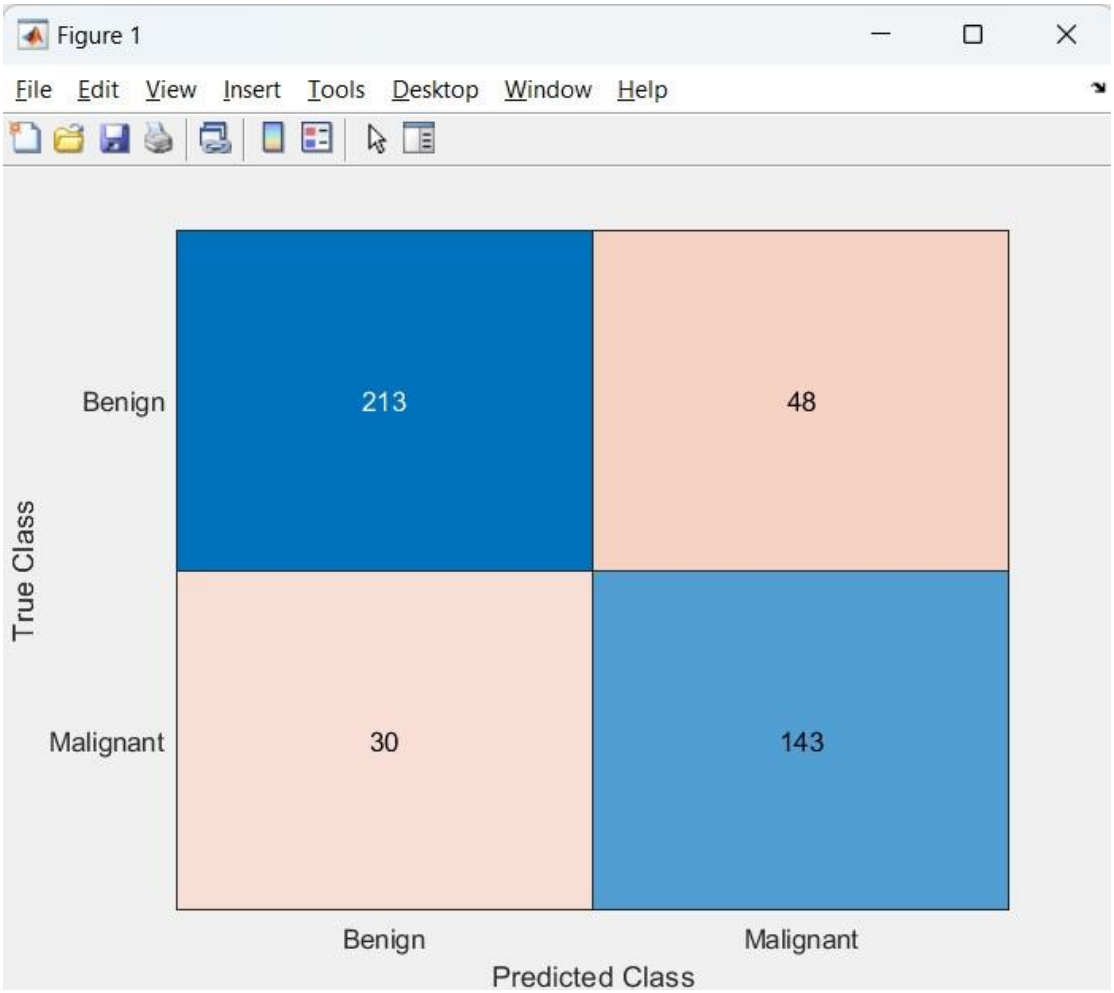

Figure 5. Confusion Matrix for Dataset 1 Using SVM
accuracy =82.02%
overall precision =81.26%
overall recall =82.13%
fl_score=81.69%

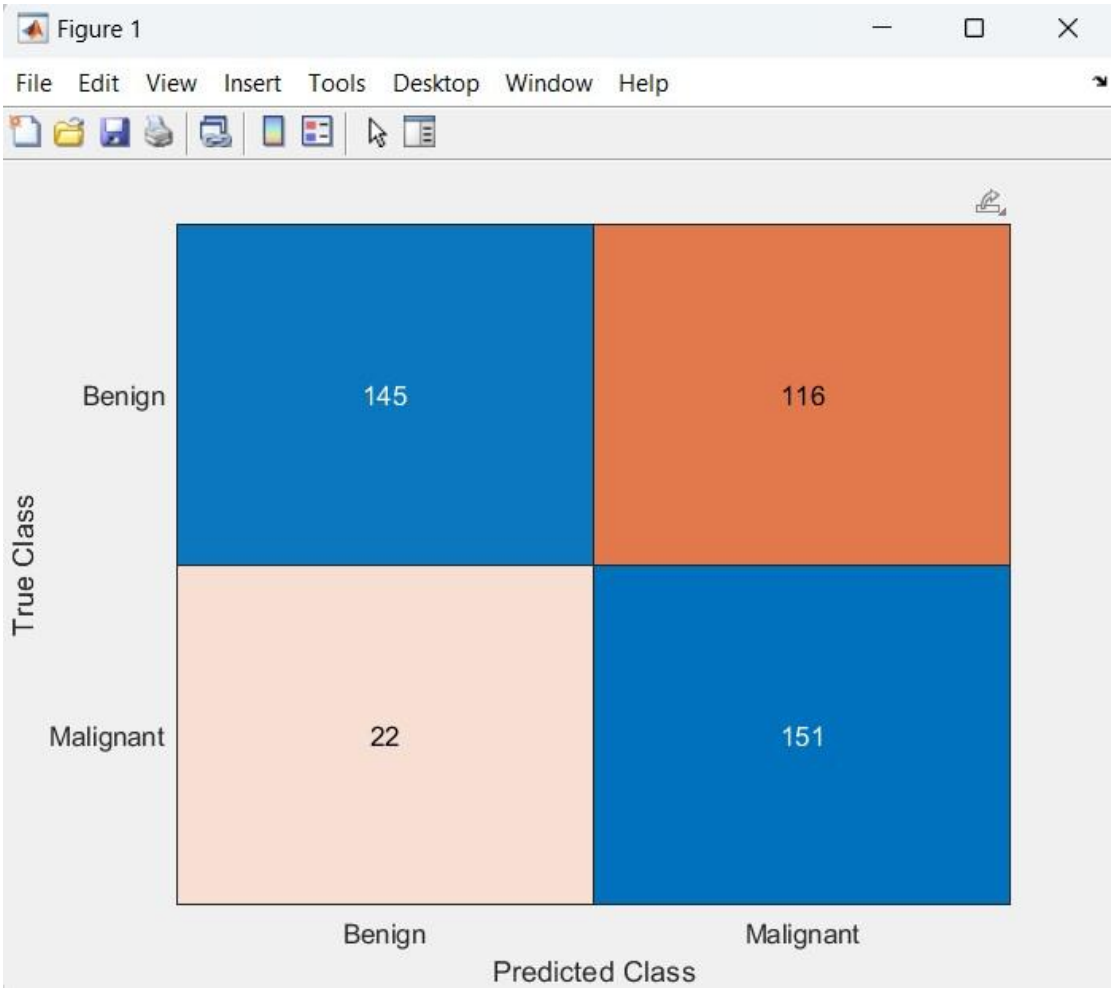

Figure 6. Confusion Matrix for Dataset 1 Using Decision Tree
accuracy =68.20%
overall precision=71.69%
overall recall =71.41%
f1_score=71.55%

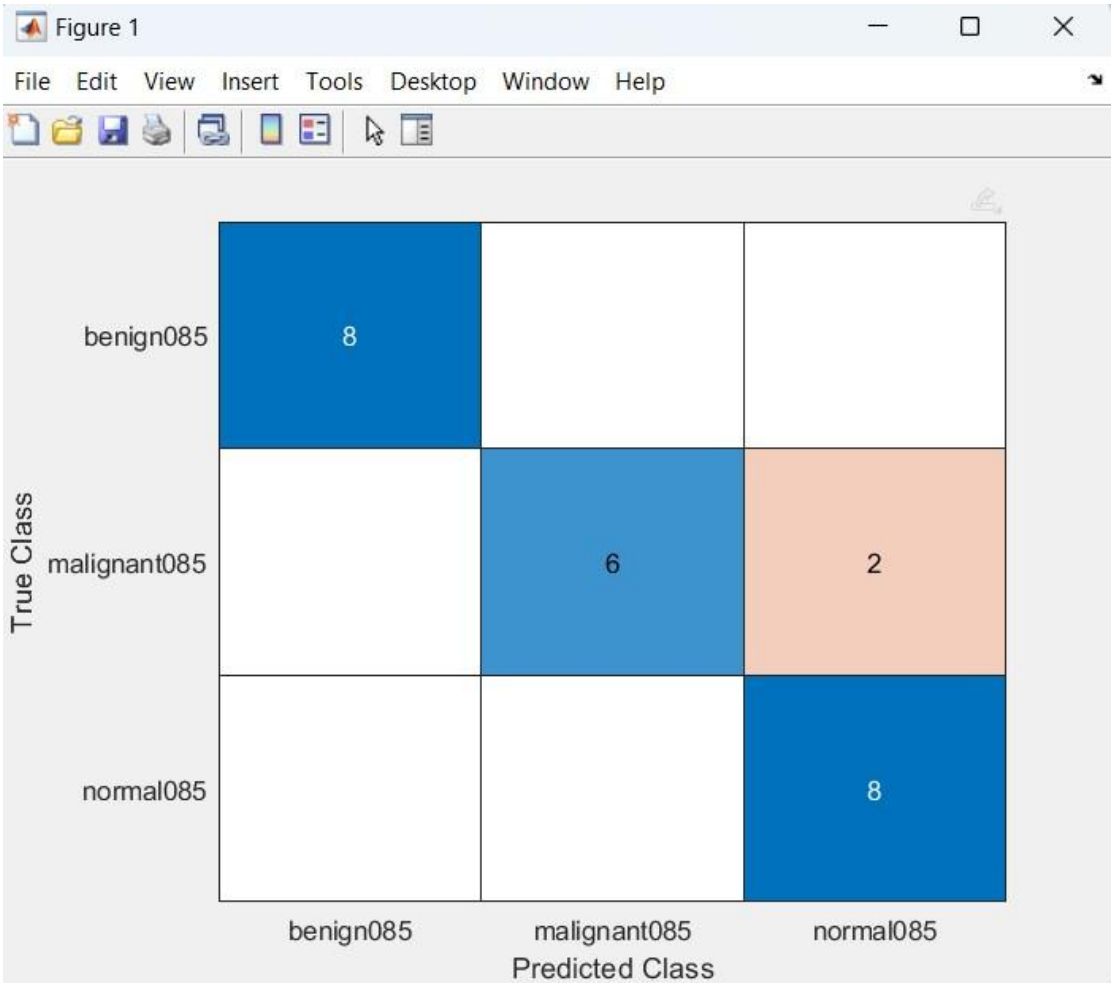

Figure 7. Confusion Matrix for Dataset 2 Using SVM

accuracy=91.66%
overall precision=93.33%
overall recall=91.66%
f1_score=92.49%

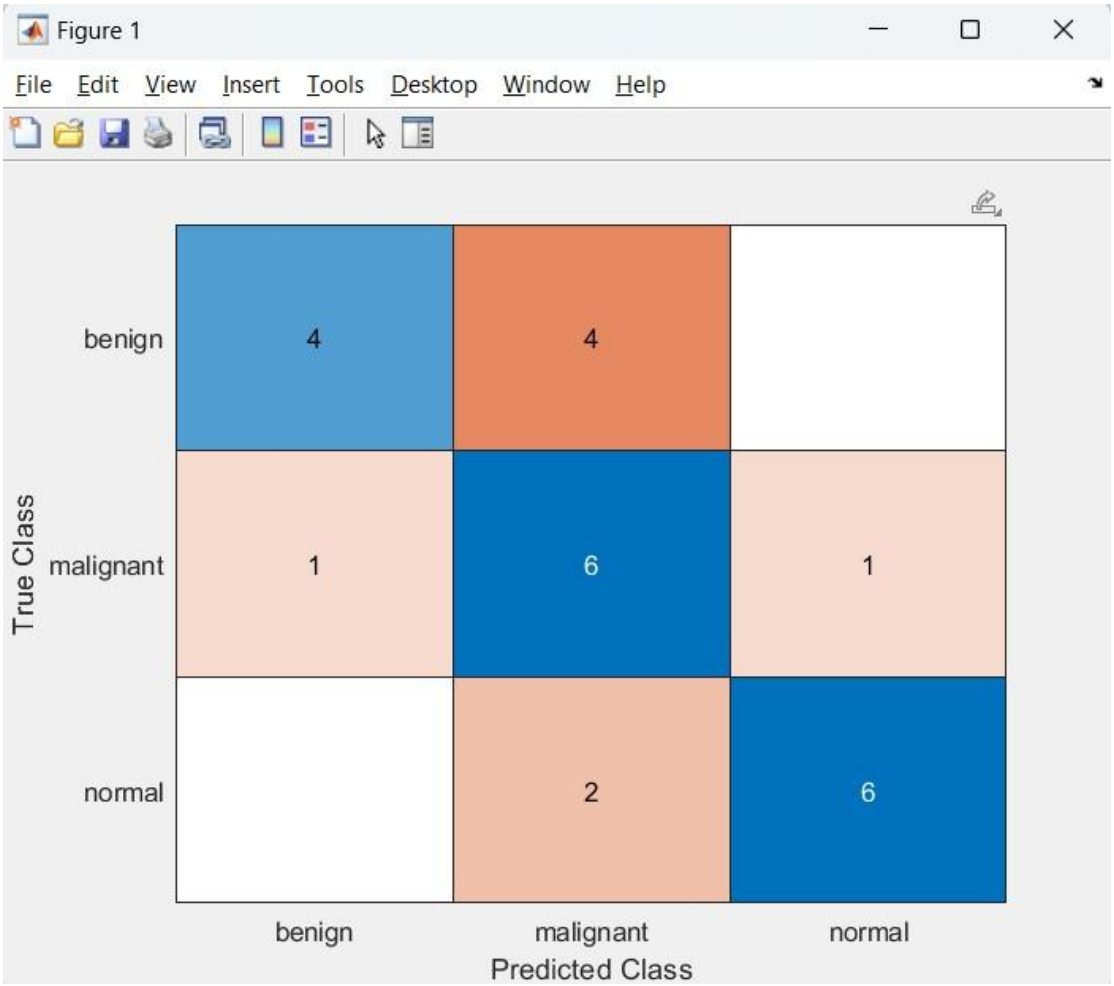

Figure 8. Confusion Matrix for Dataset 2 Using Decision Tree
accuracy =66.66%
overall precision =71.90%
overall recall =66.66%
fl_score=69.18%

TABLE 2 Comparison Table for Dataset1

S. No	PERFORMANCE METRIC	SVM	DECISION TREE
1.	ACCURACY	82.02%	68.20%
2.	PRECISION	81.26%	71.69%
3.	RECALL	82.13%	71.41%
4.	F1SCORE	81.69%	71.55%

Table-2 illustrates the Comparison of Performance Metrics of SVM And the Decision Tree for Dataset1

TABLE 3: Comparison Table for Dataset2

S. No	PERFORMANCE METRIC	SVM	DECISION TREE
1.	ACCURACY	91.66%	66.66%
2.	PRECISION	93.33%	71.90%
3.	RECALL	91.66%	66.66%
4.	F1SCORE	92.49%	69.18%

Table-3 illustrates the Comparison of Performance Metrics of SVM And Decision Tree for Dataset2

VI. CONCLUSION

The proposed system utilizes two classification algorithms, a decision tree and a support vector machine (SVM), for mammogram images analyzation. The system has an accuracy of 68.2% and 82.02% and a precision of 71.69% and 81.26% for dataset1 when utilizing the decision tree and SVM algorithms, respectively. Similarly, for dataset 2, the system has an accuracy of 66.66% and 91.66% and a precision of 68.61% and 93.33% when using the decision tree and SVM algorithms, respectively.
According to these findings, the SVM algorithm outperforms the decision tree algorithm in terms of accuracy and precision.

REFERENCES

[1] Vaishnavi Patil, Shravani Burud, Goutami Pawar, Tanaya Rayajadhav, & Sunil B. Hebbale. (2020). Breast Cancer Detection using MATLAB Functions. Advancement in Image Processing and Pattern Recognition, 3(2), 1-6.

[2] Vishal Deshwal and Mukta Sharma. Breast Cancer Detection using SVM Classifier with Grid Search Technique. International Journal of Computer Applications 178(31):18-23, July 2019. [3] Y.Ireaneus Anna Rejani et al /International Journal on Computer Science and Engineering Vol.1(3), 2009, 127-130.

[4] Zhiqiong Wang, Mo Li, Huaxia Wang, Hanyu Jiang, Yudong Yao, Hao Zhang, And Junchang Xin “Breast Cancer Detection using Extreme learning MachineBased on Feature Fusion with CNN Deep Features”, IEEE Access, August 14, 2019.

[5] Jing Zheng, Denan Lin, Zhongjun Gao, Shuang Wang, Mingjie He, And Jipeng Fan “Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis”, IEEE Access, June 4, 2020.

[6] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm, " 2016 9th International Conference on Developments in Systems Engineering (DeSE), Liverpool, 2016, P. 35-39.

[7] Saad Awadh Alanazi, M. M. Kamruzzaman, Md Nazirul Islam Sarker, Madallah Alruwaili, Yousef Alhwaiti, Nasser Alshammari, and Muhammad Hameed Siddiqi “Boosting Breast Cancer Detection Using Convolutional Neural Network, ” Journal of Healthcare Engineering Volume 2021, 5 April 2021.