International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 69-74, April-2024

https://doi.org/10.47001/IRJIET/2024.804009

Developing a Model to Identify and Analyze Features in Mammography Images to Detect Breast Cancers

[1]Hayder Raheem Salman AL-Hraishawi, [2]Ali H. Hamie

[1,2]Computer Science Department, American University of Culture & Education, Beirut, Lebanon

***Abstract* - Worldwide, breast cancer is the primary cause of mortality associated with cancer in females. Swift detection, classification, and assessment of this neoplasm can greatly diminish the corresponding fatality rate. Physical examinations have been supplanted by digital mammography as the prevailing technique for identifying breast cancer. Machine learning can use medical files and imagery to improve the early identification of conditions, optimize remedy consequences. The accuracy of determining whether or not the person with most cancers or no cancer based totally at the kind of approach which utilized for prognosis. Therefore, this study at built a convolutional neural network to extract characteristics from the DDSM mammography dataset, which have been trained and tested with several machine learning algorithms. The system achieved a detection accuracy of as much as 94% for breast cancer the usage of numerous categorization algorithms. This outcome holds good sized significance and practicality in enhancing the control of this disease and advancing its identification.**

Keywords: Breast Cancer, Convolutional Neural Network, Identify Tumor, Machine Learning, Mammography.

I. INTRODUCTION

Breast cancer is the most commonplace type of most cancers international, exceeding lung cancer with 2.3 million instances [1]. In 2020, 25% of worldwide most cancers instances had been because of breast most cancers and it is growing in many places, specifically in developing international locations. Mammography is chosen for breast cancer screening because of its top-notch sensitivity and accuracy [2]. It is an X-ray examination of the breast that detects adjustments in tissue and determines the scale of the tumor and the density of breast tissue. Despite advances, these detection strategies are not without some drawbacks as they rely on human interpretation, which can be subjective and result in different diagnoses throughout radiologists in addition to trouble in detecting early abnormalities and subtle cancers. Manual interpretation and analysis also take time in high-extent environments, which delays prognosis.

Machine learning in breast cancer is under development and can improve early detection, treatment outcomes, and patient care using vast amounts of data in this field [3]. However, the accuracy and reliability of these models must be verified and balanced between automation and human understanding when making therapeutic choices. The main focus of this study is to leverage the ability of convolutional neural networks to extract features from mammograms and use machine learning classifiers to identify cancer and enhance diagnostic accuracy.

II. RELATED WORK

The detection and study of lesions are often performed in research on healthy tissues and cancerous lesions, and mammography is the most researched imaging technique due to its widespread use. In [4], three supervised predictive models were tested to classify mammograms as normal, benign, or malignant. Random Forest excels in enhanced images, raw image classification, and multi-class labeling. Noise and artifacts were removed from the input mammography images during preprocessing. Then, feature extraction and dimensionality reduction helped reduce the amount of classification features, and analysis of variance was used to identify the most prominent features while researchers in [5] used filtering operations during image enhancement to eliminate noise and normalize the image, and threshold-based segmentation was used.

Researchers in [6] proposed the use of CAD system and SVM technology and according to the research [7], 135° is the ideal viewing angle to distinguish abnormal tissue from normal tissue and k-NN was the most effective classification method to identify malignant tumors.

Researchers in [8] sought a breakthrough in deep learning methods to distinguish between benign mammograms from negative and cancerous screenings. The CNN models were trained and tested using 14,860 images from 3,715 patients and the resulting AUC ranged from 0.76 to 0.91, while the researchers in [9] divided the procedure into three steps: firstly, processing breast images to reduce noise and improve contrast, secondly, the location and size of the tumor based on pixel density, thirdly, determining the sphericity of the tumor. In [10] a reliable detection method was presented where features were extracted by image processing and then DWT identified the affected areas of the tumor while several

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 69-74, April-2024

https://doi.org/10.47001/IRJIET/2024.804009

convolutional neural networks were tested in [11] where the MIAS dataset containing 322 mammograms was used. The DenseNet network in [12] has been used for feature extraction and a multi-view feature fusion network model for mammography classification from two perspectives. Two mammograms from different angles were used in the algorithm, and two feature branches were extracted. Some research has used transfer learning as in [13] where a third cycle of transfer learning was proposed to create a "two-view classifier" that uses caudal and medio lateral mammograms.

In [14], two new feature extraction algorithms using stepwise logistic regression were presented where the combination of mini-MIAS and DDSM datasets gave an accuracy of 85.42%. In [15], researchers recommend using DE-Ada, a basic but accurate approach to breast mass classification. First, the complementary parameters are extracted and then the mid-level features are combined using the dynamic weight of any feature or cross-modal satisfactory semantics. Finally, two voting-based ensemble learning algorithms enable delayed feature fusion. The researchers in [16] processed GAN samples to include or exclude cancerous tissue in which 1375 x 750 DDSM thumbnail images were analyzed and the classification network compared the datasets with and without these modifications, correcting for missing data, to determine the expected accuracy.

III. METHODOLOGY

In this section, a convolutional neural network will be built to extract basic features from mammograms. In addition, multiple machine learning classifiers will be used to improve the accuracy of tumor detection and evaluate overall outcomes. Figure 1 shows the basic steps of the proposed methodology.

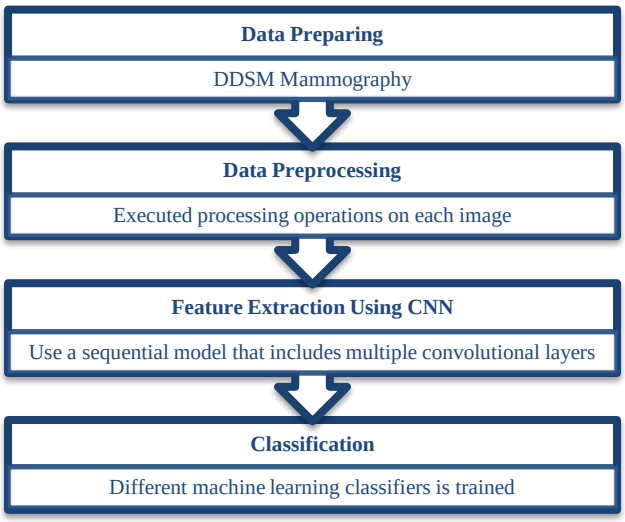

Figure 1: Proposed methodology for identifying cancer on mammograms

The Python language was used and various libraries were imported for deep learning, data processing, and image processing, such as:

- "numpy" for numerical operations and matrix processing.
- "Pandas" for processing, analyzing data, and simplifying data pre-processing.
- "TensorFlow" for deep learning.
- "OpenCV" for image analysis and processing in computer vision applications.
- "PIL" for image processing, including reading, and writing.
- "matplotlib.pyplot" for plotting and visualization.

3.1 Data Preparing

The dataset combines positive images of the CBIS-DDSM dataset and negative images of the DDSM dataset. Information was prepared by resizing the original files to 299 × 299. Masks were used to extract regions of interest (ROIs) from positive images (CBISDDSM), with some padding added for clarity. The format of the information in TFRecord files is defined by a feature dictionary. The dictionary defines connections between feature names and appropriate data types and formats. The data was decoded and the analyzed data was returned in Python dictionary form.

3.2 Data Preprocessing

A set of instructions was repeated across the images in the analyzed dataset. Each inserted image has been efficiently processed according to the following stages:

- Image data was taken from the analyzed dataset and saved as bytes.
- After being loaded into TensorFlow, the image bytes are decoded to uint8 type.
- The image is resized to 299 x 299 pixels.
- NumPy matrix is created from image tensor A to facilitate further processing.
- Using the OpenCV library, the image is reduced to its final dimensions of 100 x 100 pixels.
- Add the processed image to the "Images" list, which serves as a repository for the processed image data.
- The "label_normal" property is taken from the dataset and attached to the "labels" list. This list is responsible for storing labels corresponding to the processed images.

For testing, 20% of the data is chosen, while 80% is used for training. The data is then resampled to incorporate the channel dimension, which is set to 1 in this case. Grayscale images usually contain only one channel, unlike color images which often have three channels. Figure 2 displays a subset of

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 69-74, April-2024

https://doi.org/10.47001/IRJIET/2024.804009

images from the training data that have undergone processing. This provides a visual representation of the data's characteristics.

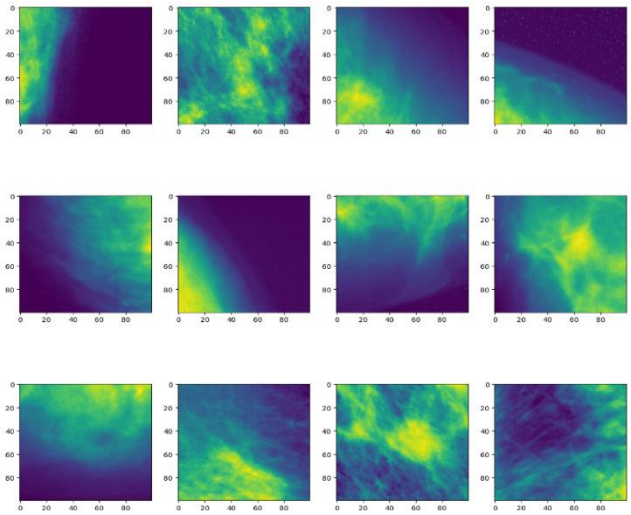

Figure 2: Images from the dataset after processing

3.3 Feature Extraction Using CNN Preprocessing

Convolutional neural networks (CNNs) are the best deep-learning model for supervised image categorization. Multiple layers are stacked to make CNNs. CNN layers are wider, taller, and deeper than normal neural network layers. This allows layers to share 215 weights according to depth [17]. These systems are effective for visual perception, pattern recognition, and picture categorization due to their order and levels. Figure 3 shows the basic CNN components: input, convolutional layers, pooling layers, fully connected (dense) layers, and output [18]. CNNs can autonomously extract hierarchical features from input data, which is crucial in computer vision applications that identify patterns and objects in images.

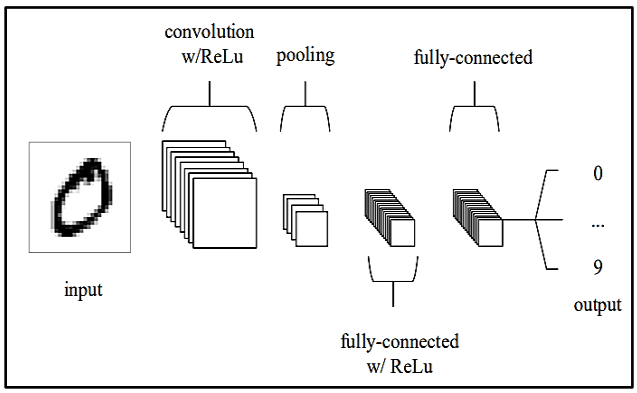

Figure 3: Basic CNN architecture [18]

The model structure is constructed sequentially as shown in Figure 4, making layer stacking easy. The architecture contains Convolutional layers, activation functions, max-pooling layers, dropout for regularization, and fully connected layers comprise.

Layer (type)	Output Shape	Param
conv2d_28 (Conv2D)	(None, 1000, 1000, 32)	288032
activation_31 (Activation)	(None, 1000, 1000, 32)	0
max_pooling2d_28 (MaxPooling2D)	(None, 500, 500, 32)	0
conv2d_29 (Conv2D)	(None, 500, 500, 64)	18496
activation_32 (Activation)	(None, 500, 500, 64)	0
max_pooling2d_29 (MaxPooling2D)	(None, 250, 250, 64)	0
dropout_24 (Dropout)	(None, 250, 250, 64)	0
conv2d_30 (Conv2D)	(None, 250, 250, 128)	73856
activation_33 (Activation)	(None, 250, 250, 128)	0
max_pooling2d_30 (MaxPooling2D)	(None, 125, 125, 128)	0
dropout_25 (Dropout)	(None, 125, 125, 128)	0
conv2d_31 (Conv2D)	(None, 125, 125, 256)	295168
activation_34 (Activation)	(None, 125, 125, 256)	0
max_pooling2d_31 (MaxPooling2D)	(None, 63, 63, 256)	0
dropout_26 (Dropout)	(None, 63, 63, 256)	0
conv2d_32 (Conv2D)	(None, 63, 63, 128)	295040
activation_35 (Activation)	(None, 63, 63, 128)	0
max_pooling2d_32 (MaxPooling2D)	(None, 32, 32, 128)	0
dropout_27 (Dropout)	(None, 32, 32, 128)	0
conv2d_33 (Conv2D)	(None, 32, 32, 64)	73792
activation_36 (Activation)	(None, 32, 32, 64)	0
max_pooling2d_33 (MaxPooling2D)	(None, 16, 16, 64)	0
dropout_28 (Dropout)	(None, 16, 16, 64)	0
conv2d_34 (Conv2D)	(None, 16, 16, 32)	18464
activation_37 (Activation)	(None, 16, 16, 32)	0
max_pooling2d_34 (MaxPooling2D)	(None, 8, 8, 32)	0
dropout_29 (Dropout)	(None, 8, 8, 32)	0
flatten_4 (Flatten)	(None, 2048)	0
dense_4 (Dense)	(None, 1)	2049
activation_38 (Activation)	(None, 1)	0

Figure 4: Proposed CNN model layers

- Convolutional Layers: The network commences with a convolutional layer of 32 filters, each having dimensions of (3, 3). The layer employs 'same' padding and strides of (1, 1). The subsequent convolutional layers have a pattern of progressively rising and then lowering the quantity of filters (64, 128, 256, 128, 64, 32).
- Activation and Pooling Layers: The incorporation of non-linearity is accomplished by employing a Rectified Linear Unit (ReLU) activation function directly following each convolutional layer. Max-pooling layers with a pool size of (2, 2), 2 strides, and 'same' padding are used to do down sampling on the spatial dimensions.
- Dropout Layers: Following each and every other max-pooling layer, dropout layers with a rate of 0.2 are inserted to reduce the likelihood of overfitting the model.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 69-74, April-2024

https://doi.org/10.47001/IRJIET/2024.804009

- Flatten Layer: The last max-pooling layer is followed by a flattened layer, which is responsible for converting the two-dimensional feature maps into a one-dimensional array. It is required to complete this stage to get the data ready for the fully linked layers.
- Fully Connected Layers: To implement a feature extraction layer, a dense layer that is comprised of 32 neurons and employs the ReLU activation function is utilized. A solitary neuron with a sigmoid activation function makes up the final layer of the neural network, which is responsible for binary classification tasks.

This architectural design seeks to effectively capture intricate characteristics included in input photos, hence rendering it highly suitable for a wide range of classification tasks, including the discrimination between malignant and benign tumor classifications.

The early stop callback is employed to monitor the training process and halt it prematurely if specific criteria are met. It aids in avoiding an excessive variety of equipment and can streamline training time. The neural network model was assembled according to the parameters shown below and the training procedures for the model were prepared:

- The 'Adam' optimizer is a widely used method for training neural networks. It is efficient because it dynamically adjusts the learning rate during the training phase.
- The loss function chosen is binary cross-entropy. This loss function is commonly employed in binary classification problems, where the network predicts two classes.
- The model's accuracy will be measured and monitored as a performance metric during training and evaluation.

The model fit function is used in the process of training the neural network with the data and settings that are provided.

3.4 Classification

The accuracy metric is commonly employed to assess the overall validity of model predictions in a two-class classification problem [19]. The statistic computes the accuracy rate by dividing the number of correctly classified cases (including true positives and true negatives) by the total number of cases. Accuracy is a quantitative measure that evaluates the model's ability to correctly identify different situations. It is distinguished by its uncomplicated nature and straightforwardness in understanding. Based on the recovered features, the effectiveness of several different machine learning classifiers was trained and evaluated. Table 1 shows a description of the algorithms used and the accuracy that each algorithm gave in detecting the tumor.

Table 1: Overview of the used machine learning algorithms and their tumor detection accuracy

ML Algorithm	Description	Test Accuracy
Extreme Gradient Boosting	Ensemble learning approach XGBoost constructs decision tree models using gradient boosting. It is known for its predictive power.	94%
Random Forest	It is an ensemble method that combines forecasts from several decision trees using the Random Forest technique. It is used for classification and regression tasks because it is trustworthy and less likely to overfit.	94%
Support Vector Classifier	It is used for binary classification. It seeks the hyperplane that maximizes group differentiation.	93%
MLP Classifier	A feedforward neural network with several hidden layers. It can learn complex data correlations, making it helpful for huge datasets and deep learning workloads.	92%
Gradient Boosting	It is an ensemble like XGBoost. It builds decision trees sequentially, fixing defects from the previous tree. It is famous for its accurate predictions.	94%
Logistic Regression	Linear classifier for binary classification. This procedure is simple and effective. Simulates the likelihood that an input belongs to a category.	94%
K-Nearest Neighbors	It is a parameter-free classification approach. The majority class of the k nearest feature space neighbors is used to assign a class label.	93%
Decision Tree	It is a hierarchical structure where nodes reflect characteristics and branches represent decision results. Decision trees are also called decision matrices.	91%
AdaBoost	AdaBoost, or adaptive boosting, is an ensemble method that combines less accurate classifiers to get a more accurate output. It adjusts training instance weights to emphasize harder-to-categorize examples.	94%
Naive	It is based on Bayes' theorem and is a probabilistic classification	85%

ISSN (online): 2581-3048
Volume 8, Issue 4, pp 69-74, April-2024
https://doi.org/10.47001/IRJIET/2024.804009

Bayes	technique. The Gaussian variation of the model implies a normal distribution of attributes.	

IV. RESULTS AND DISCUSSIONS

A higher accuracy value indicates that the classifier makes more accurate predictions. It appears that XGBoost, Random Forest, Gradient Boosting, and Logistic Regression do well. The model's creation, application to feature extraction, and training on several classifiers improved accuracy, with the greatest accuracy reaching 94%. The results of the study will be compared with the findings of prior research on breast cancer diagnosis in mammography images, which utilized the same dataset (DDSM). The comparison will be centered on the classification accuracy measure, which is presented in Table 2.

Table 2: Comparing with related work

Ref.	Methods	Accuracy
[9]	Automatic preprocessing and Efficient SRG	91.4%
[12]	CNN	92.24%
[13]	Transfer Learning	85.13%
[14]	Neighbor Structural Similarity	85.42%.
[15]	AdaBoost	90.91%.
[16]	GAN	89.6%
Proposed System	CNN (Feature Selection) + Different machine learning classifiers	94%

The proposed system shows a good level of performance in identifying breast cancer, with an accuracy rate of up to 94%. This exceeds the accuracy achieved by other previous methodologies that used a similar dataset. This indicates the importance of using deep learning, especially CNN networks, in extracting features from images to achieve a valuable contribution in the field of breast cancer identification using mammography images.

V. CONCLUSION

Early detection of breast cancer speeds up and reduces treatment procedures and saves lives. In this study, a convolutional neural network is designed for mammography classification from the DDSM mammography dataset. The images were processed and scaled to meet the requirements of a deep learning model and used to extract features from images. Subsequently, these features were fed into several machine learning algorithms. The system detected breast cancer with up to 94% accuracy using most classifiers. This study contributes to the diagnosis of breast cancer and is important in the field of health care and scientific research.

REFERENCES

[1] H. Sung, J. Ferlay, RL. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209-249, 2021.

[2] C. Canelo-Aybar, DS. Ferreira, M. Ballesteros, M. Posso, N. Montero, I. Solà, and Z. Saz-Parkinson, "Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer." *Journal of medical screening*, vol. 28, no. 4, pp. 389-404, 2021.

[3] P. Ferroni, FM. Zanzotto, S.Riondino, N. Scarpato, F. Guadagni, and M. Roselli, "Breast cancer prognosis using a machine learning approach," *Cancers*, vol. 11, no. 3, pp. 328, 2019.

[4] V. H. Viswanath, L. Guachi-Guachi, and S. P. Thirumuruganandham, "EasyChair Preprint Breast Cancer Detection Using Image Processing Techniques and Classification Algorithms Breast Cancer Detection Using Image Processing Techniques and Classification Algorithms," *EasyChair*, pp. 1-11, 2019.

[5] S. Ara, A. Das, and A. Dey, "Malignant and benign breast cancer classification using machine learning algorithms," *2021 International Conference on Artificial Intelligence (ICAI)*. IEEE, pp. 97-101, 2021.

[6] F. A. K. Al-Fahaidy, B. Al-Fuhaidi, I. Al-Darouby, F. Al-Abady, M. Al-Qadry, and A. Al-Gamal, "A diagnostic model of breast cancer based on digital mammogram images using machine learning techniques," *Applied Computational Intelligence and Soft Computing 2022*, 2022.

[7] M. Y. Kamil and A. L. A. Jassam, "Analysis of tissue abnormality in mammography images using gray level co-occurrence matrix method," *Journal of Physics: Conference Series*, vol. 1530, no. 1, 2020.

[8] S. S. Aboutalib, A. A. Mohamed, W. A. Berg, M. L. Zuley, J. H. Sumkin, and S. Wu, "Deep learning to distinguish recalled but benign mammography images in breast cancer screening," *Clinical Cancer Research*, vol. 24, no. 23, pp. 5902–5909, Dec. 2018.

[9] N. Shrivastava and J. Bharti, "Breast tumor detection and classification based on density," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 26467–26487, 2020.

ISSN (online): 2581-3048
Volume 8, Issue 4, pp 69-74, April-2024
https://doi.org/10.47001/IRJIET/2024.804009

[10] S. Kaymak, A. Helwan, and D. Uzun, "Breast cancer image classification using artificial neural networks," *Procedia Computer Science*, pp. 126–131, 2017.

[11] Y. Jin and Y. Zheng, "Medical Image Processing with Deep Learning: Mammogram Classification and Automatic Lesion Detection," *Comput. Sci. Med*, pp. 1-19, 2019.

[12] C. Zhang, J. Zhao, J. Niu, and D. Li, "New convolutional neural network model for screening and diagnosis of mammograms," *PLoS One*, vol. 15, no. 8, 2020.

[13] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, "Breast Cancer Diagnosis in Two-View Mammography Using End-to-End Trained Efficient Net-Based Convolutional Network," *IEEE Access*, vol. 10, pp. 77723–77731, 2022.

[14] R. Rabidas, A. Midya, and J. Chakraborty, "Neighborhood structural similarity mapping for the classification of masses in mammograms," *IEEE J Biomed Health Inform*, vol. 22, no. 3, pp. 826–834, 2018.

[15] H. Zhang, R. Wu, T. Yuan, Z. Jiang, S. Huang, J. Wu, J. Hua, Z. Niu, and D. Ji, "DE-Ada*: A novel model for breast mass classification using cross-modal pathological semantic mining and organic integration of multi-feature fusions," *Information Sciences*, vol. 539, pp. 461-486, 2020.

[16] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional Infilling GANs for Data Augmentation in Mammogram Classification," *Medical image analysis*, 2018.

[17] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, "Deep learning in mammography and breast histology, an overview and future trends," *Medical image analysis*, vol. 47, pp. 45–67, 2018.

[18] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," 2015.

[19] S. Liu, F. Roemer, Y. Ge, E. J. Bedrick, Z. M. Li, A. Guermazi, L. Sharma, C. Eaton, M. C. Hochberg, D. J.Hunter, M. C. Nevitt, W. Wirth, C. K. Kwoh, and X. Sun "Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies," *Osteoarthritis Cartilage*, vol. 31, no. 9, pp. 1242–1248, 2023.

Citation of this Article:

Hayder Raheem Salman AL-Hraishawi, Ali H. Hamie, "Developing a Model to Identify and Analyze Features in Mammography Images to Detect Breast Cancers" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 4, pp 69-74, April 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.804009

****

