Image Preprocessing, Segmentation, and Classification Techniques for Detection of Breast Cancer using Mammographic Images

Richa Sharma[1,*], Amit Kamra[2], Resham Chandak[3], Archana V[1], and Abhinav Kumar[4,5]

ARTICLE INFO

Article history:
Received: 15 October 2023
Revised: 1 March 2024
Accepted: 12 April 2024
Online: 31 March 2025

Keywords:
ROI
CBIS-DDSM
Breast histopathology images
Median
CLAHE
CNN

ABSTRACT

Mammography uses two X-rays with modest doses devices for screening to acquire forms of breast cancer. Field of healthcare diagnostics uses automated mammography image segmentation as the most significant method. For healthcare diagnosis, accurate segmentation of mammography is essential for the identification of any strangeness, such as lesion tissues. In this study, we have worked on single enhancement, segmentation and classification using CBIS-DDSM. For multiple images, we have done only classification. We have also found several malignant or benign images with the help of CBIS-DDSM. We have used Breast Histopathology Images Dataset for finding number of cancerous and non-cancerous images. We have performed image enhancement with the help of median, sharpening and CLAHE filters. Image segmentation has been performed using watershed and Canny edge segmentation, and also performed image classification using CNN. We have also calculated the accuracy of the Breast Histopathology model, and also for a single image classification using CBIS-DDMS. These results show that automatic deep learning techniques are easily trained to achieve high accuracy on a variety of mammography platforms, and they demonstrate the significant potential of such techniques for the creation of clinical tools that would lessen the incidence of positive and negative screening results.

1. INTRODUCTION

According to the World Health Organisation (WHO), breast cancer is the most common disease in women globally. One of many other serious problems affected the women worldwide is breast cancer [1]. Due to cancer diagnosis in advanced stage, the early mortality rate is rising in emerging nations. Mammograms are useful tools for spotting various abnormality in the budding stage, which improves the chances of saving people with cancer in breast. To find any suspicious mass, every middle-aged woman should have a mammography once a year. Clinical examination, mammography, and ultrasound are frequently utilised techniques for cancer screening [2]. Mammography is one of them, and it is regarded for the standard of detecting cancer in breast. Despite its advantages, screened mammography carried a substantial risk of both inaccurate positive and negative results [3]. Mammography uses a specific kind of high contrast imaging, high resolution, and lower dose to identify the condition. Numerous instances, noise and lower contrast in-between diseased and any tissues make it difficult to see how precisely these images were segmented. Image segmentation is essential in computer imaging to accurately identify the size and shape of lesions present [1]. It is highly encouraged that radiologists use CAD systems to help them identify and delineate the margins of breast tumours. According to the lower signal to noise ratios and tumour diversity in terms of sizes, shape, appearances, textures, and the locations, breast tumour segmentation and classification remain challenging.

Many studies have recently been proposing to be improved in breast masses classifications performance using deep representations of breast pictures and feature combining [4]. In this research paper, we have discussed about single image enhancement, segmentation and classification using CBIS-DDSM. We are using median blur, sharpening and clahe filters for enhancement. We have applied Canny edge detection on all filters for noise removal. For segmentation, we have used Canny edge segmentation and watershed segmentation. For classification, we have used CNN which uses multiple hidden layers. We have also performed data extraction, data

[1] *School of Computer Science and Engineering, Lovely Professional University Phagwara, Punjab, India.*
[2] *Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, India.*
[3] *Computer Science with AI, Indian Institute of Information Technology, Vadodara (Gandhinagar Campus), Gandhinagar, Gujarat, India.*
[4] *Department of Mechanical Engineering and Renewable Energy, Technical Engineering College, The Islamic University, Najaf, Iraq.*
[5] *Department of Mechanical Engineering, Karpagam Academy of Higher Education, Coimbatore, India. 641021.*
**Corresponding author:* Richa Sharma; Email: richa.18364@gmail.com, pdfdrabhinav@gmail.com.

cleaning and visualization using CBIS-DDSM. In this visualization, we have displayed bar diagrams for full mammograms, cropped and ROI mask images. We have also displayed pathology bar diagram. In this paper, we are also classifying 4000 images of Histopathology images dataset into cancerous and non-cancerous images using CNN classifier. We have visualized 8 cancerous and 8 non-cancerous images from these 4000 histopathology images. We are taking 1000 images for testing and 3000 images for training which are being used to calculate testing and training accuracy. We are making efforts in preprocessing techniques like ROI extraction and image enhancement to improve accuracy. After preprocessing, we have made effort in segmenting cancer images which are recognizing substantial hurdles. Segmenting an image mostly serves to streamline it and make analysis easier. More precisely, it is the process of assigning labels to each pixel in a picture so that pixel with the similar labels have specific visuals properties [5]. An accurized segment is usually the one where a pixel from a similar category has comparable multivariate value in the greyscale and create a connecting regions. Comparable values are found in neighboring pixels that belong to distinct categories. In the background removal algorithm using the morphological feature was employed as the segmentation strategy in this study. After segmenting image, we have applied CNN to categorize into cancerous and non-cancerous images. CNN is a convolutional neural network which uses multiple hidden layers like convolution, max pooling, dense and flatten. Hidden layers take in a set of weighted inputs and produce an output through an activation function. Activation functions like ‘relu’ and ‘softmax’ are being used for determining how a node responds to its inputs. Generally, neural network uses backpropagation in which known inputs are mapped with desired output. In our research, we have used only 1 classifier, that is, CNN, to classify both the segmented images which are being pre-processed using several filters and also calculated accuracies for both the segmented images. We have also used histopathology images to classify cancer and non-cancer images. In histopathology, there are only cancer and non-cancer images and not having any ‘.csv’ file. I have calculated both training and testing accuracy for histopathology images. Both the accuracies were lesser as compared to all the images of histopathology images as we have taken less training and testing images. We have taken only few images because more images will take more time to compute. Even though the particular goal might not have connected to the early trained datasets, the model weights parameter is initialized to identify basic feature like edge, corner, and texture, which will be easily utilized for not so similar tasks. This frequently reduces trained times and boosts the models accuracy.

The breast cancer detection research has made considerable strides, but there is always potential for improvement. Future study will concentrate on contrast enhancement, the incorporation of improved feature extraction methods, and the investigation of alternative classifiers such as Support Vector Machines (SVM) and Random Forest. To improve training and testing accuracy, the histopathology dataset will be enlarged. To achieve the best balance of model complexity and efficiency, fine-tuning and optimization strategies will be investigated. The study adds to our understanding of breast cancer detection and paves the way for future research, emphasizing the importance of embracing new techniques and diversifying methodologies in developing more accurate and efficient diagnostic tools in the ongoing battle against breast cancer.

2. METHODOLOGY

In this research, we have performed image enhancement, segmentation and classification on mammographic images from CBIS-DDSM database using python. In this paper, we have discussed both single and multiple images classification using CNN. For single image, we have performed image enhancement, segmentation and classification using CBIS-DDSM. We have also used CBIS-DDSM to find number of malignant, benign, full mammograms and cropped images. For multiple images, we have performed CNN classification using Breast histopathology images. We have also found the number of cancer and non-cancer images present in the Breast histopathology images.

2.1 Database

Breast Histopathology Images and the 5 GB CBIS-DDSM dataset have been used. The Curated breast imaging subset of DDSM (CBIS DDSM) is a standardized and modified version of the Digital database for screening mammography (DDSM). Breast Histopathology Images database included 162 mount slide pictures of specimens of breast cancer that were 40x scanned. 277,524 no. of patches of 50*50 size were obtained from this.

2.2 For a single image

Here, we are using CBIS-DDSM.

2.2.1 Image enhancement

We are enhancing the image to remove Gaussian noise and bring more clarity. We have used median filter, sharpening filter and CLAHE for image enhancement. We have also detected noise with the help of Canny Edge detection.

a) Median Blur Filter: We are applying median filter to remove Gaussian noise which means salt and pepper noise. Here, we are using OpenCV module for applying median filter. We use the median filter because we believe it may effectively reduce the noise caused by salt and pepper that is commonly seen in mammography images.

b) Sharpening Filter: Median filter reduces noise, but due to blurring image, image loses weight. To add weight to the image, we are sharpening the image by increasing the size of the kernel. The need to restore image weight that has been compromised by the blurring effect is what motivates the usage of a sharpening filter, with the kernel size adjustment supporting particular image properties.

c) Contrast limited adaptive histogram equalization (CLAHE): CLAHE is a contrast enhancement technique. Initially, it was employed to improve low-contrast medical photographs. CLAHE differs from regular AHE in that it reduces contrast. CLAHE was chosen because of its capacity to improve contrast, which is commonly lacking in mammographic pictures.

d) Canny Edge Detection is used for detecting noise. It is used for every filter to check its noise. It accepts input as grayscale image and employs a multistage algorithm. The use of Canny Edge Detection for noise detection demonstrates our commitment to testing each filter's performance in reducing undesirable artifacts.

2.2.2 Image segmentation

The process of segmenting an image into various areas or segments, each of which corresponds to a different object or portion of the image, is known as image segmentation. In the context of breast cancer detection, we do image segmentation to separate the breast tissue from the background, identify the boundaries of the breast, and then isolate suspicious areas, such as masses or calcifications, for the further analysis. There are various techniques for image segmentation, including thresholding, region growing, edge detection, and clustering. In the case of breast cancer detection, we use two methods. They are:

a) Watershed method: Watershed segmentation is a method used to segment an image. It involves converting the image to grayscale, thresholding, morphological opening, dilation, distance transform, foreground region, 8-bit image, unknown region, connected components identified, marker values set to zero, and watershed algorithm applied to markers. The result is displayed with red boundaries around the segmented regions. The watershed method was chosen because of its capacity to successfully segment images based on gradient information, resulting in discrete borders for detailed examination. This method is ideal for highlighting areas of interest in breast photos with diverse textures and structures.

b) Canny edge method: The OpenCV library is used to detect edge and contour detection on an input image. Gaussian blur is applied to reduce noise, the cv2.Canny function is used to detect edges, the cv2.findContours function is used to find contours, and the cv2.drawContours function is used to draw contours. The resulting image is displayed using the cv2.imshow function. Algorithm 4 shows Canny edge segmentation. The Canny edge approach was chosen for its ability to detect edges and accurately identify borders, which is crucial in isolating suspicious regions for precise investigation. This method is useful when small details in breast photos need to be accurately identified.

2.2.3 Image classification

Classification is the process of predicting class of the given input points of segmented image. We have used multilayered convolutional neural network (CNN) for classification. CNN is a deep learning algorithm which is designed for tasks where recognition of objects is very important like image detection, segmentation and classification. We have done classification for both the segmented images. Our choice of a CNN is based on its demonstrated efficacy for complex image recognition tasks. CNNs' hierarchical feature learning capabilities fits the intricate nature of picture classification perfectly. We thoroughly documented the CNN model architecture in our proposed methodology, including the type of model, layer configurations, and activation functions used. These architectural details are critical in enabling the network to capture and learn characteristics from segmented images. The careful selection of activation functions, such as 'relu' and 'softmax', improves the network's capacity to respond to inputs, which has a significant impact on the overall performance of the picture classification process.

2.3 For multiple images

We are using CBIS-DDSM and Breast Histopathology Images datasets for multiple images. We have done study for CBIS-DDSM and Breast Histopathology images separately.

2.3.1 CBIS-DDSM

We have done data extraction, data cleaning and visualization using CBIS-DDSM. We have found several images for full mammogram, cropped and ROI mask images which are based on bar diagram. We have found classification and mass images. We have also found pathology like malignant and benign images in both mass and calcification training sets. We paid close attention to any variances in imaging methods and data sources to ensure our findings represent real-world settings. We employed stratified sampling to ensure that our dataset accurately represented varied situations, avoiding biases in our research.

2.3.2 Breast histopathology images

We have used this dataset to classify 4000 images. We have divided dataset into training and testing images. Training data is used to train the algorithm. We have split 4000 images into 3000 and 1000 images for training and testing data respectively. We have found cancerous and non-cancerous images from this dataset. We have also found the model accuracy. It is critical to recognize potential differences in image quality and pathology representation within the Breast Histopathology Images dataset. Our technique entailed taking these variances into account throughout the training phase, boosting the model's adaptability to various variables seen in real-world scenarios.

2.4 Flowchart for breast cancer detection

Figure 1.1 shows stages involved in Breast Cancer detection using Mammograms. According our flowchart, firstly we have performed ROI extraction. Then, we have performed image enhancement to enhance the image. Afterwards, we have performed 2 segmentation methods and done classification for each segmented image. On the basis of classification, we are telling whether it is cancerous or not.

2.5 Result table for breast cancer detection

Table 1 shows result table for breast cancer detection. In the Result table, we have mention various researches which are done by other authors. We have mentioned author name, dataset, techniques used, findings and accuracy output.

In realm of healthcare study many other innovative techniques in healthcare showing new ways to diagnose COVID-19 were studied. Their work includes combining different types of data, using X-ray images and cough samples to train deep learning models, and using ensemble methods. This could open up new areas for early detection and classification [24-25]. Aggarwal et al.'s study of COVID-19 risk prediction for diabetic patients using fuzzy inference system and machine learning also adds a useful factor to the area where health and technology meet [26-28].

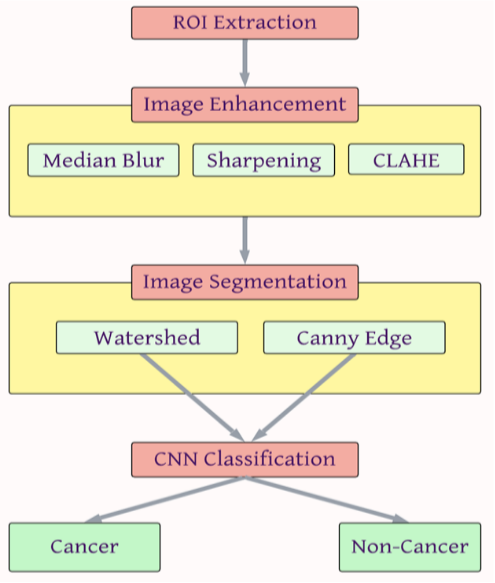


Fig. 1. Flowchart for Breast Cancer Detection using CBIS-DDSM.

Table 1. Result Table for Breast Cancer Detection

Author	Dataset	Techniques Used	Findings	Accuracy
Hiren K. Mewada et al (2020) [6]	BreaKHis Dataset and BCC Datasest	**Haar wavelet, CNN**	To enhance the performance of the classifiers, CNN model is updated and spectral signals of Haar wavelets are combined with spatial characteristics.	BreaKHis: 97.58% BCC: 97.45%,
Jyoti Parashar et al in (2020) [7]	DDSM and MIAS mammography images	K-mean cluster for segmentation. Classification: CNN-Xgboost, CNN-RNN and CNN	This study identifies breast cancer and conducts some basic research using a deep learning experience. The suggested method uses DDSM and Mias mammography pictures. In the suggested methodology, CNN hybridization is achieved by optimising tree-based learning and enhancing cumulative and class-wise performance.	CNN Xgboost: 95% CNN-RNN:96% CNN: 96.78%
Salama, Wessam M., and Moustafa H. Aly. (2021) [8]	MIAS, CBIS-DDSM and DDSM	Segmentation: U Net model Classification: Inception version 3, Dense network 121, Residual network 50, Visual geometry group 16 and Mobile network version 2	This research gave some classification methods to classify different mammographic images into benign and malignant.	98.87%

Sánchez-Cauce, Raquel, Jorge Pérez-Martín, and Manuel Luque (2021) [9]	DMR: Thermal images	Multi-input CNN	This study employs a novel strategy for early diagnosis of breast cancer by merging thermal images from multiple perspectives with personal and clinical data, resulting in the development of a multi-input classification model that takes advantage of the capabilities of CNN for image analysis.	97%
Ghosh, Swarup Kr, Anirban Mitra, and Anupam Ghosh (2021) [10]	MedPix, MIAS	Segmentation models: IFSS, IFSRCM	Breast lesions, tumors, and nodules can be more accurately partitioned using the suggested IFSRCM technique, assisting in the early detection of breast cancer. Using accuracy and roughness scores from the soft-information builder at multigranulation approximation space via defuzzification process, IFSS based clustering model IFSRCM generates a threshold picture.	JSC metric: MedPix: 93.44% MIAS: 94.06%
Heenaye-Mamode Khan, Maleika, et al. (2021) [11]	CBIS-DDSM, UPMC	Resnet 50, CNN	They have created an improved deep learning model where one of the most crucial factors in neural network training is learning rate.	88%
Sharma, Shallu, Rajesh Mehra, and Sumit Kumar (2021) [12]	BreakHis	CNN	An automated method for magnification independent multi-classification of breast cancer histopathology pictures is proposed using CNN architecture.	80.47% at 40X magnification level.
Yadav, Rohit, and Richa Sharma (2021) [13]	MIAS	Enhancement: CLAHE Feature extraction: CNN Classification: SVM and Decision tree Final decision fusion: Voting classifier	In this paper, they have detected breast cancer on the basis of decision fusion, which uses Voting classifier, of machine learning algorithms.	93.4%
Jasti, V. Durga Prasad, et al. (2022) [14]	MIAS	Image preprocessing: Geometric mean filter Feature extraction: AlexNet Feature Selection:Relief algorithm Classification: LS-SVM	To assist in the classification and detection of skin diseases, this model integrates image preprocessing, feature extraction, feature selection, and machine learning approaches.	Nearly 98%.
Rafid, AKM Rakibul Haque, et al. (2022) [15]	CBIS-DDSM database	Binary Masking, Contour Detection, Canny Edge detection, CLAHE, Random Forest, KNN, Multilayer Perceptron, Naive Bayes, SVM, RF, PCA and Wrapper Method	The study in the identification of cancer in breast from mammographic images has been compared to another research on a same dataset. This approach is most effect in assisting radiologist in many clinics insight that included detecting the cancer at an early stage, and less errors.	98.05%
Srivastav, Gaurav, et al. (2022) [16]	MIAS database	K-Means, median filter, random forest. Hybrid ML, MLP, Decision Tree, and morphological operations.	In this paper, we are developing a combination of ML and DL models training on the dataset of mammographic images that are improving the last risk of the model and obtaining the accuracy in the range of radiologist for cancer in	89.05%

			breast screening.	
Mohapatra, Subasish, et al. (2022) [17]	Mini-DDSM	CNN Architectures: AlexNet, VGG16, and ResNet5	In this study, they examine some of the CNN classifiers that are used to identify breast cancer in mammography pictures by dividing them into benign, cancerous, and normal classes.	AlexNet: 95% VGG16: 65% ResNet5:61%
Samee, Nagwan Abdel, et al. (2022) [18]	INBREAST	AlexNet, GoogleNet and VGG-16	The shallow and deep features from the INbreast mammograms are extracted using a random selection of deep learning models from AlexNet, VGG, and GoogleNet.	98.50%
Yadav, Rohit, Richa Sharma, and Pushpendra Kumar Pateriya (2022) [19]	MIAS	Contrast Enhancement: CLAHE, HE, MMBEBHE and RMSHE Feature extraction: CNN Classification: SVM, Decision tree and RF	In this paper, they have detected breast cancer on the basis of feature and decision fusion. They discovered that MMBHBHE and HE do not perform as well as CLAHE and RMSHE.	SVM: 92.30% Decision tree: 94.03, Random forest: 95.05%
Atrey, Kushangi, et al. (2023) [20]	43 mammogram images and 43 ultrasound images collected from 31 patients	CNN and LSTM	The proposed bimodal CAD algorithm using combined mammogram and ultrasound outperforms the traditional unimodal CAD systems.	Mammography : 97.16% Ultrasound: 98.84%
Majumdar, Samriddha, Payel Pramanik, and Ram Sarkar (2023) [21]	BreakHis, ICIAR-2018	CNN models: GoogleNet, VGG11 and MobileNetV3_Small	The suggested ensemble model is aimed to solve a 2-class classification issue of breast histopathology images using the Gamma function.	BreakHis: 98.67% for 200X of magnification. ICIAR-2018: 96.95%
Aniwat Juhong et al (2023) [22]	The female MUC1 doubletransgenic mice with breast tumors	CNN model, SRGAN-ResNeXt architecture, modified U-Net architecture, Generator model, Discriminator model.	Custom CNN can help overcome the inaccessibility of advanced microscopes in remote clinics in developing nations where low performance microscopes are usually found.	91.2%
Eren Tekin et al (2023) [23]	Tubule-U-Net	Residual Network34, Dense Network 161 reflection padding technique, mirror padding technique, CNN segmentation, rectified activation function for linear unit and max pool layer.	Based on results, TubuleUNetwork, that is based on Efficient Network B3 U Network, used for training patches, which is obtained using reflection padding and it's testing is obtained by using overlapped strategy, which supplies highest specificity, and false positive rates in comparison to other approaches	ResNetwork34 UNetwork: 80.57% DenseNetwork 161UNetwork: 81.73% EfficientNetworkB3UNetwork :86.33%

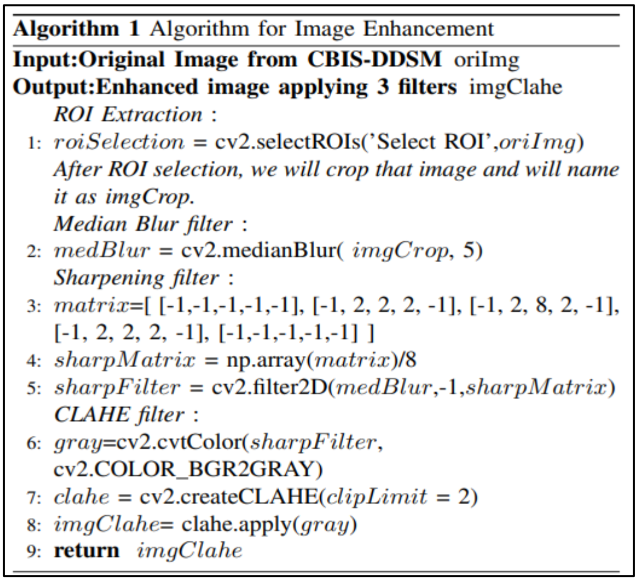
Kumar et al. break new ground in medical diagnostics with fuzzy neural techniques for postpartum hemorrhage, while Sharma et al.'s deep learning models redefine tuberculosis detection [29-32]. Their innovative approaches highlight the potential for advanced classification techniques in the study of breast cancer.

3. PROPOSED ALGORITHM

3.1 For a single image

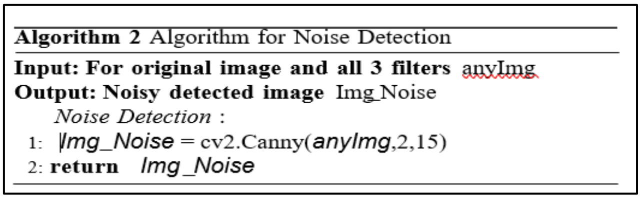
Here, we are using CBIS-DDSM.

ROI Extraction and applying image enhancement: We have used selectROIs() function for selecting ROI and after that we have cropped original image according to selected ROI. Then, we used median blur filter, sharpening filter, and contrast enhancement filter for enhancing image. We have also detected noise present in all images using Canny Edge detection. Algorithm 1 shows image enhancement function, which tells us about ROI extraction, median blur filter, sharpening filter, clahe respectively.

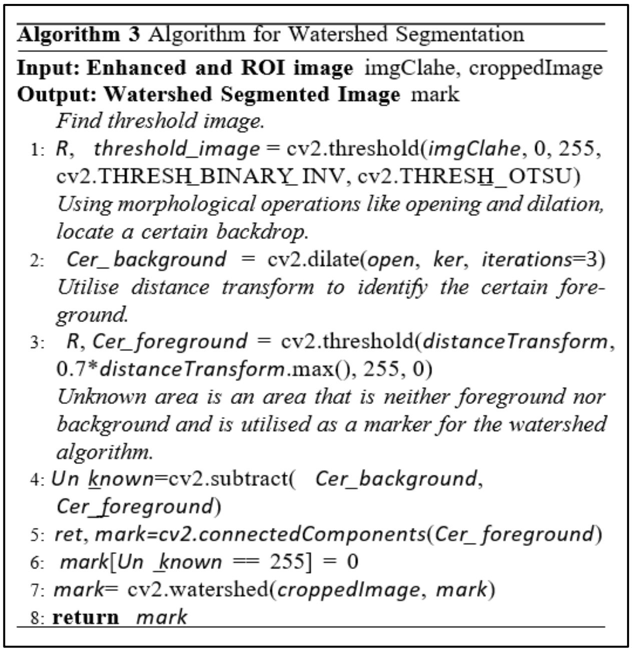


```
Algorithm 1 Algorithm for Image Enhancement
Input:Original Image from CBIS-DDSM oriImg
Output:Enhanced image applying 3 filters imgClahe
   ROI Extraction :
   roiSelection = cv2.selectROIs('Select ROI',oriImg)
   After ROI selection, we will crop that image and will name
   it as imgCrop.
   Median Blur filter :
   medBlur = cv2.medianBlur( imgCrop, 5)
   Sharpening filter :
   matrix=[ [-1,-1,-1,-1,-1], [-1, 2, 2, 2, -1], [-1, 2, 8, 2, -1],
   [-1, 2, 2, 2, -1], [-1,-1,-1,-1,-1] ]
   sharpMatrix = np.array(matrix)/8
   sharpFilter = cv2.filter2D(medBlur,-1,sharpMatrix)
   CLAHE filter :
   gray=cv2.cvtColor(sharpFilter,
   cv2.COLOR_BGR2GRAY)
   clahe = cv2.createCLAHE(clipLimit = 2)
   imgClahe= clahe.apply(gray)
   return imgClahe
```

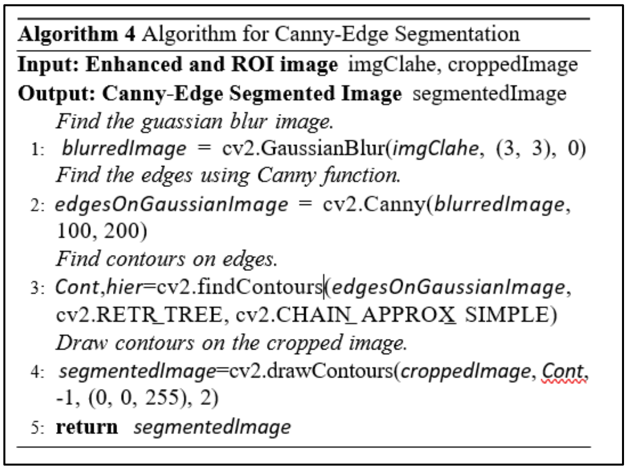
Algorithm 2 shows Noise detection using Canny-edge detection.



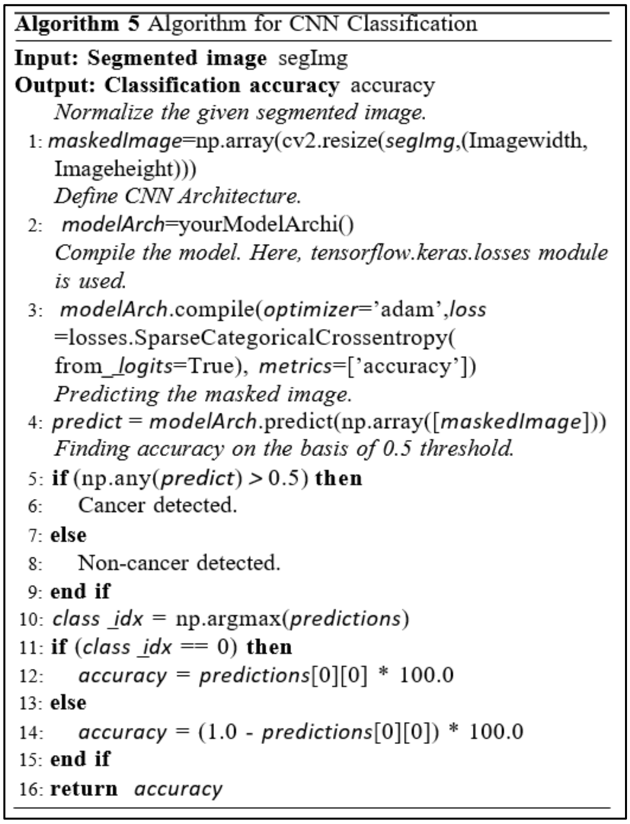
```
Algorithm 2 Algorithm for Noise Detection
Input: For original image and all 3 filters anyImg
Output: Noisy detected image Img_Noise
   Noise Detection :
   Img_Noise = cv2.Canny(anyImg,2,15)
   return Img_Noise
```



```
Algorithm 3 Algorithm for Watershed Segmentation
Input: Enhanced and ROI image imgClahe, croppedImage
Output: Watershed Segmented Image mark
   Find threshold image.
   R, threshold_image = cv2.threshold(imgClahe, 0, 255,
   cv2.THRESH_BINARY_INV, cv2.THRESH_OTSU)
   Using morphological operations like opening and dilation,
   locate a certain backdrop.
   Cer_background = cv2.dilate(open, ker, iterations=3)
   Utilise distance transform to identify the certain fore-
   ground.
   R, Cer_foreground = cv2.threshold(distanceTransform,
   0.7*distanceTransform.max(), 255, 0)
   Unknown area is an area that is neither foreground nor
   background and is utilised as a marker for the watershed
   algorithm.
   Un_known=cv2.subtract( Cer_background,
   Cer_foreground)
   ret, mark=cv2.connectedComponents(Cer_foreground)
   mark[Un_known == 255] = 0
   mark= cv2.watershed(croppedImage, mark)
   return mark
```



```
Algorithm 4 Algorithm for Canny-Edge Segmentation
Input: Enhanced and ROI image imgClahe, croppedImage
Output: Canny-Edge Segmented Image segmentedImage
   Find the guassian blur image.
   blurredImage = cv2.GaussianBlur(imgClahe, (3, 3), 0)
   Find the edges using Canny function.
   edgesOnGaussianImage = cv2.Canny(blurredImage,
   100, 200)
   Find contours on edges.
   Cont,hier=cv2.findContours(edgesOnGaussianImage,
   cv2.RETR_TREE, cv2.CHAIN_APPROX_SIMPLE)
   Draw contours on the cropped image.
   segmentedImage=cv2.drawContours(croppedImage, Cont,
   -1, (0, 0, 255), 2)
   return segmentedImage
```



```
Algorithm 5 Algorithm for CNN Classification
Input: Segmented image segImg
Output: Classification accuracy accuracy
   Normalize the given segmented image.
   maskedImage=np.array(cv2.resize(segImg,(Imagewidth,
   Imageheight)))
   Define CNN Architecture.
   modelArch=yourModelArchi()
   Compile the model. Here, tensorflow.keras.losses module
   is used.
   modelArch.compile(optimizer='adam',loss
   =losses.SparseCategoricalCrossentropy(
   from_logits=True), metrics=['accuracy'])
   Predicting the masked image.
   predict = modelArch.predict(np.array([maskedImage]))
   Finding accuracy on the basis of 0.5 threshold.
   if (np.any(predict) > 0.5) then
       Cancer detected.
   else
       Non-cancer detected.
   end if
   class_idx = np.argmax(predictions)
   if (class_idx == 0) then
       accuracy = predictions[0][0] * 100.0
   else
       accuracy = (1.0 - predictions[0][0]) * 100.0
   end if
   return accuracy
```

3.2 Image segmentation

We have segmented enhanced images with 2 different methods- Watershed and Canny edge segmentation methods.

3.2.1 Watershed segmentation algorithm

Algorithm 3 shows Watershed Segmentation function.

3.2.2 Canny edge segmentation algorithm

Algorithm 4 shows Canny edge segmentation function.

3.3 Image classification

Algorithm 5 shows CNN classification. In our algorithm, we classified both the segmented images. This algorithm contains 3 steps:

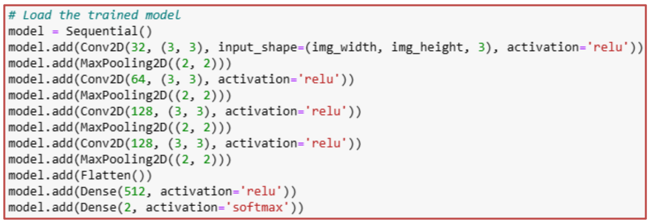
a) Normalize the given segmented image.

b) Define CNN architecture.

c) Predicting the masked image.

d) Finding accuracy on the basis of 0.5 threshold. If any of the predictions are greater than 0.5, then the given image is cancerous, otherwise non-cancerous [32-35].

CNN model Architecture: In our research, we have used CNN for image classification. CNN is a concept of neural network which contain input, hidden and output layers. There are multiple hidden layers in CNN. In our code, we have used Sequential model with the help of ‘keras’. In sequential model, we have added layers like Conv2D, maxpooling2D, dense and flatten layers. For Conv2D and dense layers, we have used ‘relu’ as an activation function [36-41]. We have used the kernel size as 3x3. At the last, we have used ‘softmax’ as an activation function in the dense layer. Figure 1.2 shows CNN model architecture. It shows Sequential model and different keras layers which makes the architecture effective.



```
# Load the trained model
model = Sequential()
model.add(Conv2D(32, (3, 3), input_shape=(img_width, img_height, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(128, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(128, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Flatten())
model.add(Dense(512, activation='relu'))
model.add(Dense(2, activation='softmax'))
```

Fig. 1.2. CNN model architecture.

3.4 For multiple images

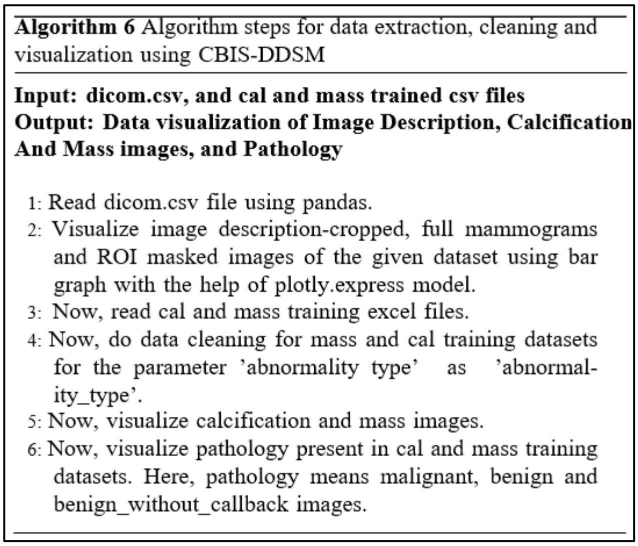
We are using CBIS-DDSM and Breast Histopathology Images datasets for multiple images. We have done study for CBIS-DDSM and Breast Histopathology images separately.

3.4.1 CBIS-DDSM

We have done data extraction, data cleaning and visualization using CBIS-DDSM. We have found several images for full mammogram, cropped and ROI mask images which are based on bar diagram. We have found classification and mass images. We have also found pathology like malignant and benign images in both mass and calcification training sets. Algorithm 6 shows steps for data extraction, data cleaning and data visualization.

3.4.2 Breast histopathology images

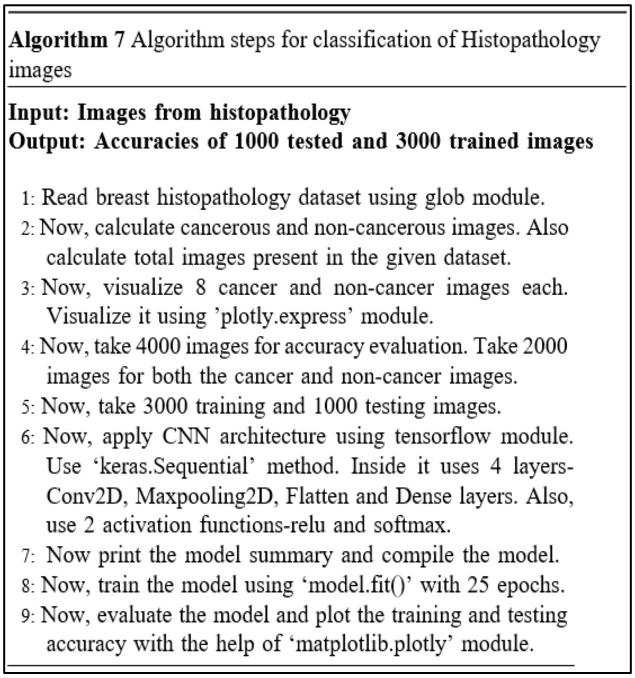
We have taken 1000 testing and 3000 training images to calculate testing and training accuracies by implementing CNN. Algorithm 7 shows CNN classification histopatholgy images.



Algorithm 6 Algorithm steps for data extraction, cleaning and visualization using CBIS-DDSM

Input: dicom.csv, and cal and mass trained csv files
Output: Data visualization of Image Description, Calcification And Mass images, and Pathology

```
1: Read dicom.csv file using pandas.
2: Visualize image description-cropped, full mammograms
   and ROI masked images of the given dataset using bar
   graph with the help of plotly.express model.
3: Now, read cal and mass training excel files.
4: Now, do data cleaning for mass and cal training datasets
   for the parameter ’abnormality type’ as ’abnormal-
   ity_type’.
5: Now, visualize calcification and mass images.
6: Now, visualize pathology present in cal and mass training
   datasets. Here, pathology means malignant, benign and
   benign_without_callback images.
```



Algorithm 7 Algorithm steps for classification of Histopathology images

Input: Images from histopathology
Output: Accuracies of 1000 tested and 3000 trained images

```
1: Read breast histopathology dataset using glob module.
2: Now, calculate cancerous and non-cancerous images. Also
   calculate total images present in the given dataset.
3: Now, visualize 8 cancer and non-cancer images each.
   Visualize it using ’plotly.express’ module.
4: Now, take 4000 images for accuracy evaluation. Take 2000
   images for both the cancer and non-cancer images.
5: Now, take 3000 training and 1000 testing images.
6: Now, apply CNN architecture using tensorflow module.
   Use ‘keras.Sequential’ method. Inside it uses 4 layers-
   Conv2D, Maxpooling2D, Flatten and Dense layers. Also,
   use 2 activation functions-relu and softmax.
7: Now print the model summary and compile the model.
8: Now, train the model using ‘model.fit()’ with 25 epochs.
9: Now, evaluate the model and plot the training and testing
   accuracy with the help of ‘matplotlib.plotly’ module.
```

4. RESULT AND DISCUSSION

We have performed single image enhancement, segmentation and classification using CBIS-DDSM dataset. Figure 1.3 (A) shows ROI extracted and enhanced images and Figure 1.3 (B) shows image segmentation through water method and canny edge method.

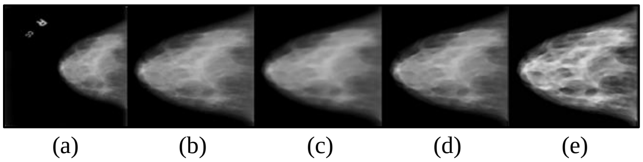
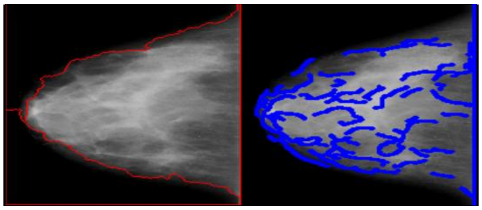


Fig. 1.3 (A). Image Enhancement Output (a) Original image (b) Roi extracted image (c) Median blur filter (d)Sharpening filter (e) Contrast enhanced filter.



(a) (b)

Fig. 1.3 (B). Image segmentation output (a) Watershed (b) Canny edge method.

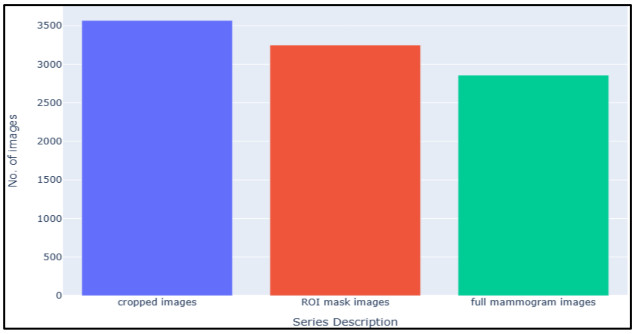


Fig. 1.4 (A) Image description (a) Blue – cropped images (b) Orange - full mammogram images (c) Green -ROI mask images.

Output for classification of watershed segmented image was found to be **cancerous with 99.67%** accuracy and its predicted class was 0. Output for classification of Canny edge segmented image was found to be **cancerous with 97.82%** accuracy and its predicted class was 0. We have also done data extraction, cleaning and visualization in sequence as stated by various papers [42]. Figure 1.4 (A) shows image description which shows number of full mammograms, cropped and ROI mask images. Figure 1.4 (B) shows a bar diagram which shows several images for abnormalities in cancer. Figure 1.4 (C) shows a bar diagram for pathology in calcification training set. Figure 1.4 (D) shows a bar diagram for pathology in mass training set.

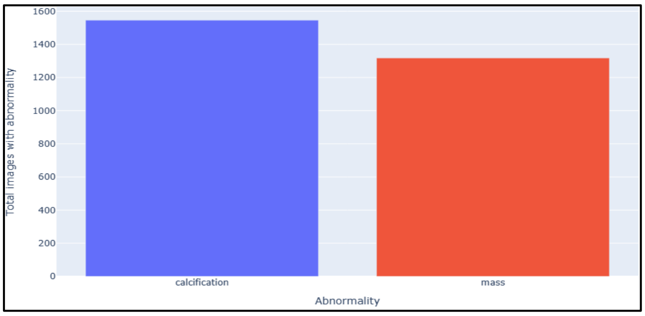


Fig. 1.4 (B) Abnormalities in cancer (a) Blue – calcification (b) Orange – mass.

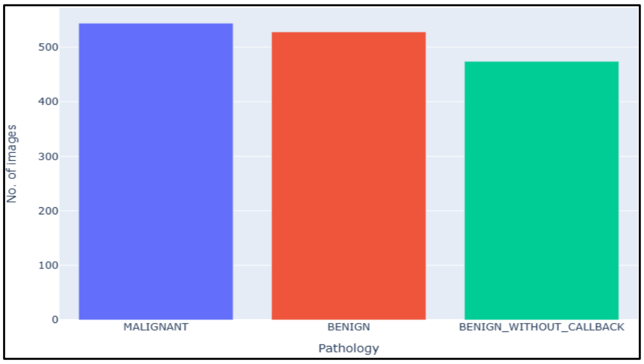


Fig. 1.4 (C) Pathology in calcification training set (a) Blue – malignant (b) Orange - benign (c) Green – benign_without_callback.

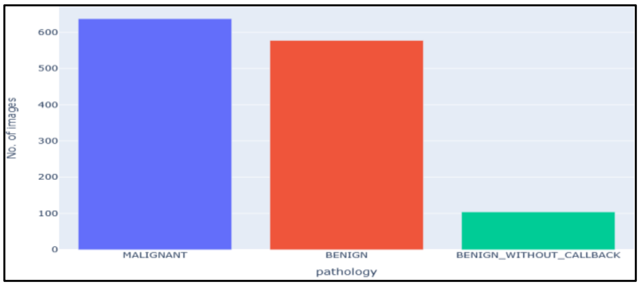


Fig. 1.4 (D) Pathology in mass training set (a) Blue – malignant (b) Orange - benign (c) Green – benign_without_callback.

We have also performed CNN classification using Breast histopathology images. We have calculated accuracy based on 4000 images of Breast histopathology Images. Accuracy output for tested images, which includes 1000 images, is 92.80% and for trained images, which includes 3000 images, is 98%. Figure 1.5 (A) shows cancer and non-cancer images. Figure 1.5 (B) shows model accuracy.

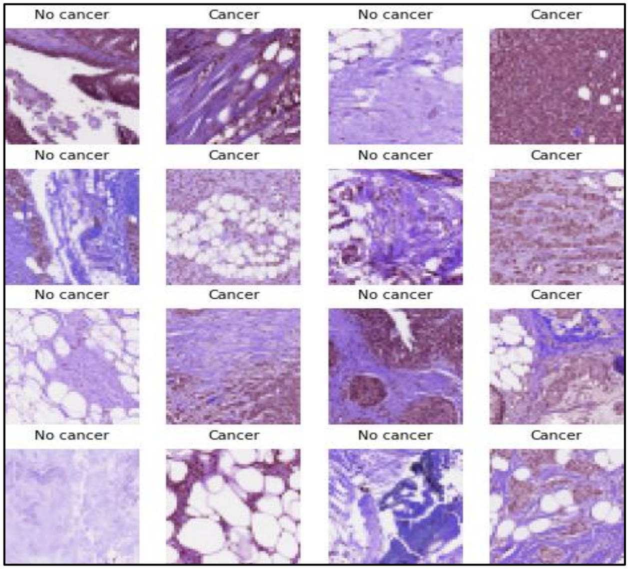


Fig. 1.5 (A). Cancer and non-cancer images.

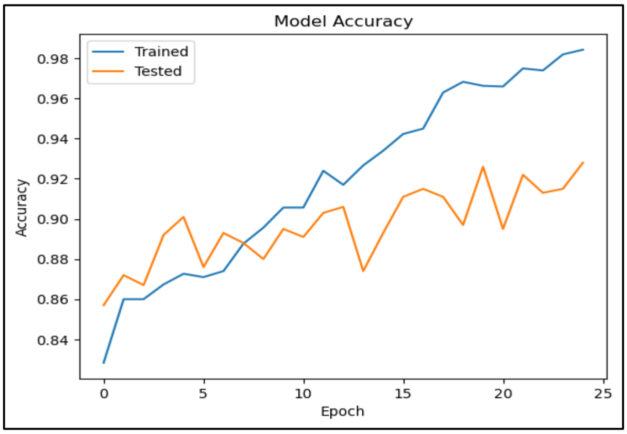


Fig. 1.5 (B) Accuracy for Breast histopathology images.

5. CONCLUSION

In this study, we presented a modified CNN model-based approach for classifying cancerous histopathology images. The standard CNN model's flaw is that it only uses spatial characteristics that can be extracted from the training dataset for classification. It looks at numerous photos from the CBIS-DDSM lawsuit against CNN to make a breast cancer diagnosis. We have also performed single image enhancement, segmentation and classification using CBIS-DDSM dataset. For single image classification, we are getting accuracies as 99.67% and 97.82% for watershed and canny edge segmented images. Here, histopathology photos were used to classify cancerous and non-cancerous images, and the results showed tested data accuracy as 92.80% for 1000 images and trained data accuracy as 98% for 3000 images. In future, we will try to improve our single image classification technique.

6. FUTURE WORK

Our future research will focus on refining and expanding our breast cancer detection technique, with the following specific goals in mind:

- **Assessment of Contrast Enhancement approaches**: We intend to investigate further contrast enhancement approaches, such as histogram equalization and adaptive histogram equalization, in order to determine the best ways for enhancing mammographic pictures.
- **Feature Enhancement Strategies**: To improve the discriminative features used in our classification model, we will integrate advanced feature extraction methods like as PCA, LDA, and RFE.
- **Classifier Diversity**: In addition to the present CNN, we will analyze the effectiveness of alternative classifiers such as SVM, Random Forest, and Decision Trees in breast cancer classification.
- **Model Optimization**: To improve model performance, the CNN architecture will be fine-tuned by modifying parameters such as layer configurations, activation functions, and regularization approaches.
- **Ethical and Bias implications**: As part of our continuous study, we will thoroughly examine ethical implications and potential biases in the models. This evaluation is critical for ensuring fair and unbiased outcomes across a wide range of patient demographics.

This reduced future work plan keeps the focus on essential areas for improvement while summarizing our research strategy.

ABBREVIATIONS

ROI: Region of interest; CBIS-DDSM: Curated breast imaging subset of Digital Database for Screening Mammograph; HE: Histogram equalization; CLAHE: Contrast-limited adaptive histogram equalization; CNN: Convolutional Neural Network; WHO: World Health Organization; CAD: Computer Aided Design; OpenCV: Open Source Computer Vision Library; BreaKHis: Breast Cancer Histopathological Image Classification; BCC: Basal-cell carcinoma; DDSM: Digital Database for Screening Mammograph; MIAS: Mammographic Imaging Analysis Society; CNN-Xgboost: CNN and Efficient extreme gradient boosting; CNN-RNN: CNN and Recurrent neural networks; U Net: U-shaped CNN; CTLM: Computed Tomography Laser Mammography; FFDM: Full-Field Digital Mammography; DMR: **Douban Movie Review Dataset; JSC:** Jaccard Similarity Coefficient; IFSS: Intuitionistic Fuzzy sets; IFSRCM: Intuitionistic Fuzzy Soft Rough C-Means; MedPix: Medical Picture Exchange, UPMC: University of Pittsburgh Medical Center; LSTM**: Long Short-Term Memory;** AlexNet: CNN Architecture; MMBEBHE: Minimum mean brightness error bi-histogram equalization; RMSHE: Root-mean-square deviation; RCNN: Region Based CNN; SVM: Support Vector Machine; SVM-RFE: SVM-Recursive Feature Elimination; KNN: K-Nearest Neighbors; RF: Random Forest; PCA: Principal Component Analysis; ML: Machine learning; DL: Deep learning; MLP: Multi-layer perceptron; MUC1: Mucin 1; SRGAN-ResNeXt : Super Resolution Generative Adversarial Network and Residual Networks with Next.

AVAILABILITY OF DATA MATERIALS

The data described in this data note can be freely and openly accessed via 'CBIS-DDSM: Breast Cancer Image Dataset (kaggle.com)' and 'Breast Histopathology Images (kaggle.com)'.

REFERENCES

[1] Soulami, Khaoula Belhaj, et al. "Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation." *Biomedical Signal Processing and Control* 66 (2021): 102481.

[2] Pavithra, M., et al. "Prediction and classification of breast

cancer using discriminative learning models and techniques." *Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches* (2021): 241-262.

[3] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh. "Deep Learning to Improve Breast Cancer Detection on Screening Mammography." Scientific Reports volume 9, Article number: 12495 (2019)

[4] Vivek Kumar Singh, Hatem A. Rashwan, Santiago Romani, Farhan Akram, Nidhi Pandey, Md. Mostafa Kamal Sarker, Adel Saleh, Meritxell Arenas, Miguel Arquez, Domenec Puig, Jordina Torrents-Barrena. "Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network." Volume 139, January 2020, 112855

[5] H.N. Abdullah and H.K. Abduljaleel, "Deep CNN Based Skin Lesion Image Denoising and Segmentation using Active Contour Method," Engineering and Technology Journal, Vol. 37, Part A, No. 11, pp. 464-469, 2019.

[6] Hiren K Mewada, Amit V,Patel, Mahmoud Hassaballah, Monagi H. Alkinani And Keyur Mahant (2020).Spectral–Spatial Features Integrated Convolution Neural Network for Breast Cancer Classification.

[7] Parashar, J., Sumiti, & Rai, M. (2020). Breast cancer images classification by clustering of ROI and mapping of features by CNN with XGBOOST learning.

[8] Salama, Wessam M., and Moustafa H. Aly. "Deep learning in mammography images segmentation and classification: Automated CNN approach." *Alexandria Engineering Journal* 60.5 (2021): 4701-4709.

[9] Sánchez-Cauce, Raquel, Jorge Pérez-Martín, and Manuel Luque. "Multi-input convolutional neural network for breast cancer detection using thermal images and clinical data." *Computer Methods and Programs in Biomedicine* 204 (2021): 106045.

[10] Ghosh, Swarup Kr, Anirban Mitra, and Anupam Ghosh. "A novel intuitionistic fuzzy soft set entrenched mammogram segmentation under multigranulation approximation for breast cancer detection in early stages." *Expert Systems with Applications* 169 (2021): 114329.

[11] Heenaye-Mamode Khan, Maleika, et al. "Multi-class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN)." *Plos one* 16.8 (2021): e0256500.

[12] Sharma, Shallu, Rajesh Mehra, and Sumit Kumar. "Optimised CNN in conjunction with efficient pooling strategy for the multi-classification of breast cancer." *IET Image Processing* 15.4 (2021): 936-946.

[13] Yadav, Rohit, and Richa Sharma. "Breast Cancer Detection Based on Decision Fusion of Machine Learning Algorithms." Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4. Springer Singapore, 2021.

[14] Jasti, V. Durga Prasad, et al. "Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis." *Security and communication networks* 2022 (2022): 1-7.

[15] Rafid, AKM Rakibul Haque, et al. "An Effective Ensemble Machine Learning Approach to Classify Breast Cancer Based on Feature Selection and Lesion Segmentation Using Preprocessed Mammograms." *Biology* 11.11 (2022): 1654.

[16] Srivastav, Gaurav, et al. "Breast Cancer Detection in Mammogram Images using Machine Learning Methods and CLAHE Algorithm." *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2022.

[17] Mohapatra, Subasish, et al. "Evaluation of deep learning models for detecting breast cancer using histopathological mammograms Images." *Sustainable Operations and Computers* 3 (2022): 296-302.

[18] Samee, Nagwan Abdel, et al. "Deep learning cascaded feature selection framework for breast cancer classification: Hybrid CNN with univariate-based approach." *Mathematics* 10.19 (2022): 3631.

[19] Yadav, Rohit, Richa Sharma, and Pushpendra Kumar Pateriya. "Feature and Decision Fusion for Breast Cancer Detection." *Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1*. Springer Singapore, 2022.

[20] Atrey, Kushangi, et al. "Mammography and ultrasound based dual modality classification of breast cancer using a hybrid deep learning approach." *Biomedical Signal Processing and Control* 86 (2023): 104919.

[21] Majumdar, Samriddha, Payel Pramanik, and Ram Sarkar. "Gamma function based ensemble of CNN models for breast cancer detection in histopathology images." *Expert Systems with Applications* 213 (2023): 119022.

[22] Aniwat Juhong, Bo Li, Cheng-You Yao, Chia-Wei Yang, Dalen W. Agnew, Yu Leo Lei, Xuefei Huang, Wibool Piyawattanametha, and Zhen Qiu, "Superresolution and segmentation deep learning for breast cancer histopathology image analysis," Biomed. Opt. Express 14, 18-36 (2023)

[23] Tekin, E., Yazıcı, Ç., Kusetogullari, H. et al. Tubule-U-Net: a novel dataset and deep learning-based tubule segmentation framework in whole slide images of breast cancer. Sci Rep 13, 128 (2023).

[24] Kumar, Santosh, et al. "A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques." *Computer methods and programs in biomedicine* 226 (2022): 107109.

[25] Kumar, Santosh, et al. "Chest X ray and cough sample based deep learning framework for accurate diagnosis of COVID-19." *Computers and Electrical Engineering* 103 (2022): 108391.

[26] Itika Sharma, Sachin Kumar Gupta*, "Channel Tracking in IRS-based UAV Communication Systems using Federated Learning", Journal of Electrical Engineering, 74(06), pp: 521-531, 2023.

[27] Kumar, Santosh, et al. "Ensemble multimodal deep learning for early diagnosis and accurate classification of COVID-19." *Computers and Electrical Engineering* 103 (2022): 108396.

[28] Amina Khan, Sumeet Gupta, Sachin Kr Gupta*, "UAV-Enabled Disaster Management: Applications, Open Issues, and Challenges," GMSARN International Journal, 18(1), 44-53, 2024.

[29] Vijay Kumar Sharma "CNTFET circuit-based wide fan-in domino logic for low power applications", Journal of Circuits, Systems and Computers, World Scientific Publishing Company, 31(2), 2022. https://doi.org/10.1142/

S0218126622500360.

[30] Chukhu Chunka, Subhasish Banerjee, Sachin Kumar Gupta*, "A secure communication using multifactor authentication and key agreement techniques in internet of medical things for COVID-19 patients", Concurrency and Computation: Practice and Experience, Wiley, 35(7), pp: 01-22, 2023, https://doi.org/10.1002/cpe.7602.

[31] Aggarwal, Alok, et al. "COVID-19 risk prediction for diabetic patients using fuzzy inference system and machine learning approaches." Journal of Healthcare Engineering 2022 (2022).

[32] Choubey, Dilip Kumar, et al. "Soft Computing Approaches for Ovarian Cancer: A Review." GMSARN International Journal 18 (2024) 223-239.

[33] Vishwakarma, Shivam, et al. "Cancer Gene Clustering Using Computational Model." GMSARN International Journal 18 (2024) 252-257.

[34] Susmita Mondal, Mehak Shafi, Sumeet Gupta, Sachin Kr Gupta*, "Blockchain based Secure Architecture for Electronic Healthcare Record Management," GMSARN International Journal, 16 (4), 413-426, 2022.

[35] Kumar, VD Ambeth, et al. "A novel solution for finding postpartum haemorrhage using fuzzy neural techniques." *Neural Computing and Applications* (2021): 1-14.

[36] Vinay Pathak, Karan Singh, Radha Raman Chandan, Sachin Kumar Gupta*, Manoj Kumar, Shashi Bhushan, Sujith Jayaprakash, "Efficient Compression Sensing Mechanism based WBAN System", Security and Communication Networks, Hindawi, 2023, Article ID 8468745, 1-12 https://doi.org/10.1155/2023/8468745

[37] Sadat Riyaz, Vijay Kumar Sharma, "Design of reversible Feynman and double Feynman gates in quantum-dot cellular automata nanotechnology", Circuit world, Emerald Publishing Limited, 49(1), 28-37, 2023.

[38] Santosh Kumar, Mithilesh Kumar Chaube, Srinivas Naik Nenavath, Sachin Kr. Gupta*, Sumit Kumar Tetarave, "Privacy Preservation and Security Challenges: A New Frontier Multimodal Machine Learning Research", International Journal of Sensor Networks, Inderscience, 39(4), 227-245, 2022.

[39] Sachin Kr. Gupta*, "Arduino-Based Wireless Hand Gesture Controlled Robot" Applied Data Science and Smart Systems, AIP Conf. Proc. 2916, 040004 (2023), 040004-1–040004-7; https://doi.org/10.1063/5.0177522 Published by AIP Publishing. 978-0-7354-4733-2/$30.00

[40] Vinayak Sharma, Sachin Kumar Gupta, and Kaushal Kumar Shukla. "Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images." *Intelligent Medicine* (2023).

[41] Akshita Gupta, Abhinav Shukla, Rahul Yadav, Shubham Mishra, Mamoon Rashid, Sachin Kr. Gupta*, *"Real-Time Hand Gesture Replication System Using 3D Modelling Software"*, IEEE Students' Conference on Engineering & Systems (SCES), MNNIT Prayagraj, India, pp: 1-5, 10-12 July 2020, DOI**:** 10.1109/SCES50439.2020.9236756.

[42] Chanchal Kumar, Avinash Sharan Mishra, Vijay Kumar Sharma, "Leakage Power Reduction in CMOS Logic Circuits Using Stack ONOFIC Technique", IEEE Second International Conference on Intelligent Computing and Control Systems (ICICCS), 1363-1368, 2018.

[43] Richa Sharma and Amit Kamra. "Enhancing diagnosis of breast cancer through mammographic image segmentation using Fuzzy C-Means." International Journal of Sustainable Building Technology and Urban Development 14.4 (2023): 488-499. Print. doi:10.22712/susb.20230038.

[44] R. Sharma and A. Kamra, "A Review on CLAHE Based Enhancement Techniques," *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 321-325, doi: 10.1109/IC3I59117.2023.10397722.