Breast Cancer Detection System Using Convolutional Neural Network

Ranjeet Yadav[1], Saurabh Maurya[2], Shivam Sharma[3], Sumit Gaurav[4], Mr. Madhup Agrawal[5]

[1,2,3,4,5]Department of IT, Ajay Kumar Garg Engineering College Ghaziabad, Uttar Pradesh, India

Abstract

Breast cancer stands as a leading cause of cancer- related fatalities worldwide. Assessing cancer accurately through eosin-stained images remains a complex task, often resulting in discrepancies among medical professionals while reaching a con- clusive diagnosis. To streamline this intricate process, Computer-Aided Diagnosis (CAD) systems present a promising avenue, aiming to reduce costs and enhance efficiency. Traditional clas- sification methods hinge on problem-specific feature extraction, rooted in domain knowledge. However, addressing the multitude of challenges posed by these feature-centric techniques has led to the emergence of deep learning methods as viable alternatives. Here, we propose a novel approach employing Convolutional Neural Networks (CNNs) for the classification of hematoxylin and eosin-stained breast biopsy images. Our method categorizes images into four distinct groups: normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma. Additionally, it per- forms a binary classification distinguishing carcinoma from non- carcinoma cases. The meticulously designed network architecture facilitates information extraction across multiple scales, encom- passing both individual nuclei and overall tissue organization. This design choice enables seamless integration of our proposed system with wholeslide histology images. Notably, our method achieves a commendable accuracy of 77.8four-class classification and demonstrates a high sensitivity of 95.6% in identifying cancer cases.

1. INTRODUCTION

Breast cancer originates from the cells in the breast and stands as one of the most prevalent forms of cancer affecting women. It ranks second, following lung cancer, as a po- tentially lifethreatening disease among women. The disease encompasses various types that are distinguished based on the appearance of cells under a microscope. The primary types include invasive ductal carcinoma (IDC) and ductal carcinoma Identify applicable funding agency here. If none, delete this. in situ (DCIS). DCIS progresses slowly and typically doesn't significantly impact patients' daily lives, Breast Cancer Detec- tion System Using Convolutional Neural Network accounting for a smaller percentage of cases (ranging between 20affecting approximately 80patients, poses higher risks as it infiltrates the entire breast tissue.

Early detection plays a crucial role in effectively treating breast cancer. Therefore, the availability of accurate screen- ing methods is vital to identify initial symptoms. Imaging techniques like mammography, ultrasound, and thermography are commonly employed for breast cancer screening. Mam- mography stands out as a pivotal method for early detection. However, ultrasound becomes more effective in cases of denser breast tissue. Radiography may overlook smaller masses while thermography might prove more adept at diagnosing such masses compared to ultrasound.

Given the challenges inherent in interpreting medical images due to low contrast, noise, and intricacies beyond human visual perception, tools have been developed to enhance im- age processing. Artificial intelligence (AI), machine learning (ML), and convolutional neural network (CNN) technolo- gies have emerged as rapidly growing sectors in healthcare. They aim to solve complex tasks by reducing reliance on human intelligence. Within machine learning, deep learning (DL) employs artificial neural networks such as deep neural networks (DNN), recurrent neural networks (RNN), deep belief networks (DBN), and CNNs. These architectures find applications across various domains including medical image analysis, aiding in the enhancement of diagnostic accuracy for cancer detection.

Breast cancer has over 90% chances of being cured com- pletely among all other cancer types. Because cancer doesn't cause pain at early stage, it doesn't get attention until the health conditions are severe. Average age among indian women reporting cancer is 35-40s. The survival rate of patients is the percentage estimate of the patients who will survive for a given period of time after the diagnosis, for the expectancy of a normal life.Survival rate varies by the stage at which cancer is detected. According to most recent data the survival rate for the breast cancer diagnosed among women are[1]:

- 91% at 5 years after diagnosis.
- 86% after 10 years.
- 80% after 15 years.

Women and men with first degree relative (parent, child or sibling) having a history of breast cancer are most prone to the disease. Mammography is a procedure which is low dose x- ray through which we can visualize breast's internal structure. Different signal processing techniques such as ultrasound imaging, microwave imaging, wavelet transform which is the time-frequency representation of a signal using small wave- form called wavelet and curvelet transform which is derived from wavelet transform are used for breast cancer detection, its degree of localization varies with scale and produce images on different scales[3]. Other techniques like fuzzy logic and neuro-fuzzy system are also used for feature extraction for breast cancer to differentiate between abnormal and normal categories[3]. Deep learning is a machine learning technique in which a computer model performs classification tasks directly learning from text, images or sound. Models are trained on a large number of datasets and CNN architectures containing many layers. In medical imaging deep learning is used to detect cancer cells automatically. Training a deep convolution network from start is difficult because it needs large amount of data for training. One way is to fine tune an existing per- trained network. Deep learning is used in various medical fields such as bioinformatics, early diagnosis of Alzheimer's disease and molecular imaging etc. Molecular imaging is a new field that combines patientspecific and disease-specific molecular information with conventional anatomical imaging readouts[4]. A new method was proposed which was capable of analyzing multiple classes in one setting which worked on minimum prior domain knowledge and required less labeled samples for training[5]

2. Method and models

A. Datasets

The primary dataset utilized in this project originates from the Radiological Society of North America (RSNA) dataset, obtained through a recent Kaggle competition [22]. Compris- ing 54,713 images, this dataset accompanies a corresponding ground truth classification distinguishing benign from malig- nant tumors. Its significance lies in its value for researchers keen on advancing machine learning algorithms for breast cancer (BC) detection. The dataset encompasses both normal and abnormal mammograms, exhibiting a spectrum of breast densities and lesion types crucial for algorithmic develop- ment. Figure 2 illustrates exemplar images from this dataset showcasing cases of cancer and normalcy. Additionally, the Digital Database for Screening Mammography (DDSM) [24] contributes 55,890 images, with 14% depicting positive cases and the remaining 86% showing negative results. These images were divided into 598 × 598 tiles, subsequently resized to 299 × 299. A subset of this dataset, denoted as CBIS-DDSM, focuses on positive cases, featuring annotations and expertextracted regions of interest. However, in this study, the CBIS-DDSM subset remains unused, and the original DDSM dataset is employed due to the specific aim of classifying images from both normal subjects and cancer patients. Figure 1 presents sample images from this dataset illustrating cases of cancer and normalcy.

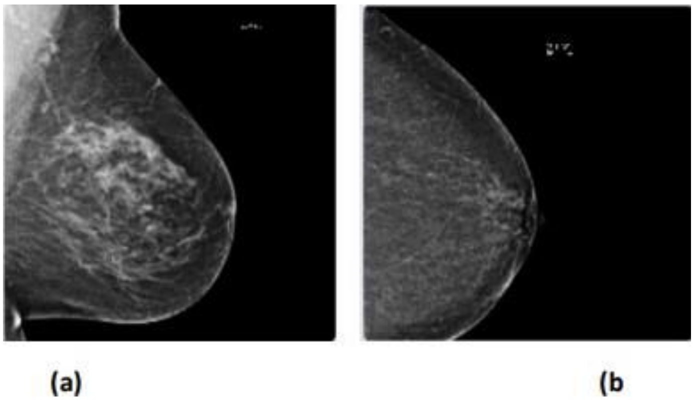


Fig. 1. -Two figure show sample images from RSNA dataset for (a) a cancerous, and (b) a normal subject.

B. The MIAS (Mammographic Image Analysis Society) dataset, referenced as [23], stands as a renowned and exten- sively employed resource for crafting and assessing Computer- Aided Detection (CAD) systems targeting breast cancer (BC) detection. Comprising 322 mammographic images, each image is complemented by an associated ground truth classification discerning benign or malignant tumors. Its value lies signifi- cantly in aiding researchers keen on advancing machine learn- ing algorithms for BC detection. This dataset encompasses a diverse array of both normal and abnormal mammograms, showcasing various breast densities and lesion types, thereby serving as a rich resource for algorithm development. Within Figure 2, two sample images representing cases of cancer and normalcy from this dataset are depicted..

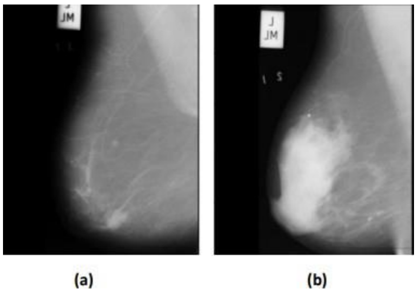


Fig. 2. These figures show two sample images from the MIAS dataset for (a) cancerous, and (b) normal subjects

C. The DDSM (Digital Database for Screening Mammog- raphy) [24] comprises a total of 55,890 images, with 14% representing positive cases and the remaining 86% depict- ing negative results. Initially organized as 598 × 598 tiles, these images were subsequently resized to 299 × 299 for standardization. A segmented subset of this dataset, known as CBIS-DDSM, specifically focusing on positive cases, has been meticulously annotated, and the regions of interest have been precisely extracted by experts. However, within the scope of this research, the CBIS-DDSM subset remains unutilized. Instead, the original DDSM dataset is employed, aligning with the study’s aim of classifying images pertaining to both normal subjects and cancer patients. Figure 3 showcases two sample images from this dataset.

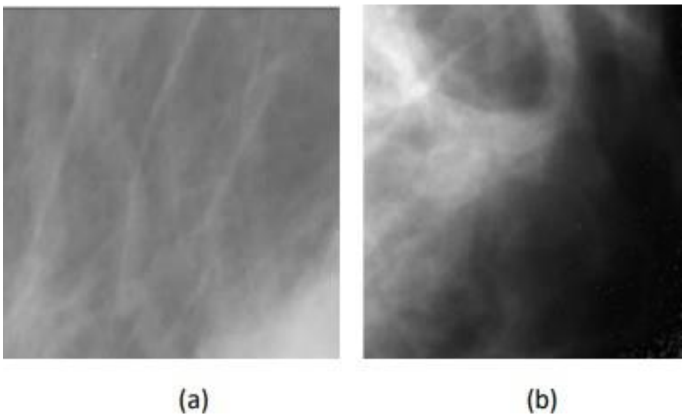
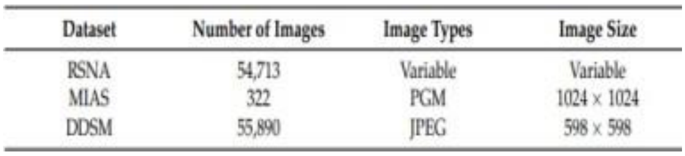


Fig. 3. These figure shows two sample images from DDSM dataset for (a) cancerous and (b) normal case



Dataset	Number of Images	Image Types	Image Size
RSNA	54,713	Variable	Variable
MIAS	322	PGM	1024 × 1024
DDSM	55,890	JPEG	598 × 598

Fig. 4. Dataset

B. 2.2 Methodology:

Training was done on 70Mammograms MIAS from a total of 322 images. Fig. 2 shows the methodology followed for the proposed system.

C. 2.3 Algorithm Used

We evaluate our approach using four different machine learning algorithms: neural network (NN), k-nearest neighbour (KNN), random forest (RF), and support vector machine (SVM). Our results demonstrate that the NN-based classifier achieves an impressive accuracy of 92% on the RSNA dataset. To train the Stochastic Gradient Descent with Momentum (SGDM), we employed various parameter adjustments such as fine-tuning the base learning rate, mini-batch size, and setting the maximum number of epochs to achieve optimal results.

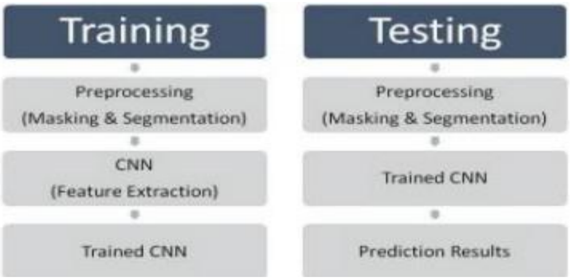


Fig. 5. Conceptual level block diagram of the training and testing procedures in the proposed system

Table II presents some of the parameters utilized during this process. Our approach in this paper involved training the Convolutional Neural Network (CNN) entirely from the initial stage, specifically tailored for our application. Within a CNN, the network's layers function akin to detection filters, designed to identify distinct patterns or features embedded within an image. Initially, the early layers of a CNN detect relatively large and easily interpret features. Subsequently, as the network progresses through its layers, it becomes adept at detecting smaller and more abstract features. Finally, the last layer of the CNN is capable of making highly refined detection's or classifications.

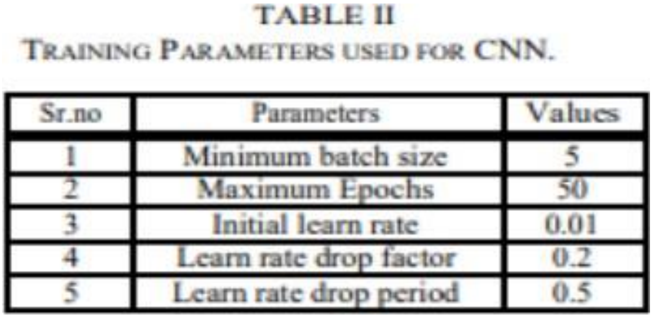


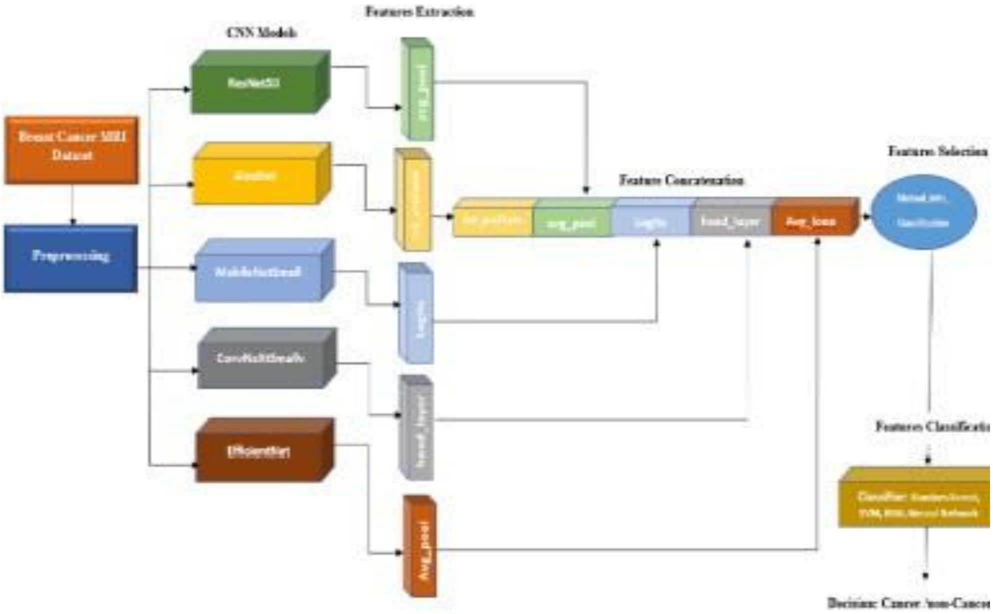
TABLE II
TRAINING PARAMETERS USED FOR CNN.

Sr.no	Parameters	Values
1	Minimum batch size	5
2	Maximum Epochs	50
3	Initial learn rate	0.01
4	Learn rate drop factor	0.2
5	Learn rate drop period	0.5

Fig. 6. Table 2

The detailed classification process involves amalgamating all the identified features from the preceding layers. The Deep Convolutional Neural Network (DCNN) comprises seven layers, as depicted in Fig. 3, where the initial four layers are convolutional layers and the remaining three layers are fully connected. The input for the DCNN consists of grayscale images.

In this architecture, each neuron conducts a computation via the dot product of weights applied to the local region con- nected within the input volume. To capture features effectively, we've employed 4, 16, and 80 filters of sizes (2, 3, 5) with padding sizes (3, 2, 1) encompassing all edges of the input layer. For instance, a filter size of [3 3] denotes filters with a



Block diagram of the proposed system.

Fig. 7.

height and width of 3 each, traversing across the input's width and height. Two pooling layers have been integrated to downsample the data, reducing computational load and bolstering the network's robustness. These pooling layers employ a filter size of 2 by 2 pixels, extracting the maximum value among four inputs within each local region.

Furthermore, the final layer typically utilizes the Soft- maxLayer in a CNN-based classifier, aiding in the probability distribution for multi-class classification tasks. The learning rate is a crucial parameter that defines the speed at which the model adapts to the data during training. changes of weights on each epoch e.g. larger learning rate determines larger weight changes on each epoch and the network learns quicker and vice versa. We have used learning rate of 0.01.

A-Training and Testing the CNN using original data The initial dataset comprised images with dimensions of 1024-by-1024. To facilitate training, we categorized the data into two distinct classes: normal and abnormal. For the training phase, we allocated 150 images for the normal class and 100 images for the abnormal class, while the remaining images were earmarked for testing purposes. Various filters of sizes 2, 3, and 5 were employed during the experiment. Addition- ally, we experimented with different methods of dividing the training and testing datasets – automatically and manually – using a 70:30 ratio, which yielded varying outcomes. Notably, training the data in a randomized manner resulted in superior performance compared to non-randomized automatic splitting. The proposed method, integrating deep learning techniques for breast cancer detection, has showcased promising potential. Fig. 6 illustrates the satisfactory outcomes achieved. This research remains ongoing, with further enhancements antic- ipated. Future efforts involve optimizing the Convolutional Neural Network (CNN) architecture and exploring the utilization of pre-trained networks, which are expected to enhance accuracy levels significantly.

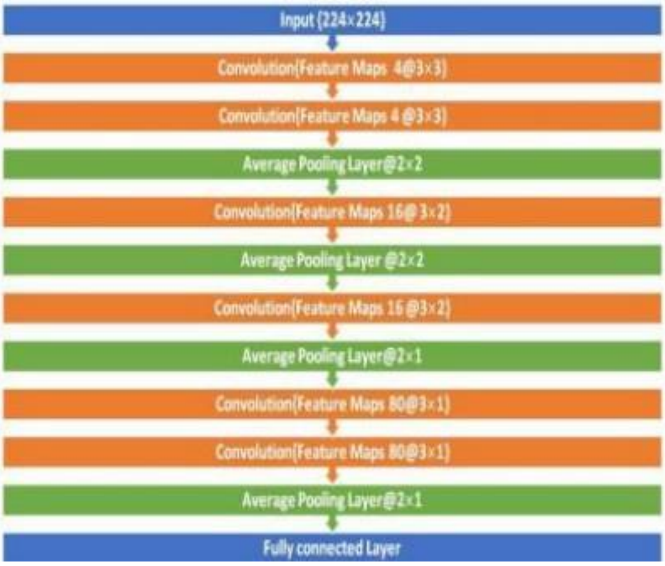

Input (224×224)
Convolution(Feature Maps 4@3×3)
Convolution(Feature Maps 4 @3×3)
Average Pooling Layer@2×2
Convolution(Feature Maps 16@ 3×2)
Average Pooling Layer @2×2
Convolution(Feature Maps 16 @3×2)
Average Pooling Layer@2×1
Convolution(Feature Maps 80@3×1)
Convolution(Feature Maps 80@3×1)
Average Pooling Layer@2×1
Fully connected Layer

Fig. 8. Archetecture of CNN of perposed System

B-Training and Testing the CNN using preprocessed data The dataset underwent preprocessing steps, where images

initially sized at 1024-by-1024 were resized to a dimension of 224-by-224. To enhance data quality, noise reduction tech- niques were applied through morphological operations such as binarization and masking, facilitating the extraction of Regions of Interest (ROIs). These operations are instrumen- tal in describing and isolating specific image components. Furthermore, the dataset was categorized into seven distinct subclasses, with six classes dedicated to various types of abnormalities and one class exclusively containing normal images. These subclasses were denoted based on the nature of abnormalities present, including architectural distortion, asymmetry, calcification, spiculated masses, circumscribed masses, and a miscellaneous category encompassing images that couldn't be conclusively identified as benign or malignant. During the experimentation phase, we conducted training and testing using the preprocessed dataset, maintaining consistency with three filter sizes (2, 3, 5). This approach allowed for a comprehensive assessment of the dataset across different filter configurations. Top of Form data with and without randomizing it. Satisfactory results were achieved on all filter sizes as shown in Fig. 7. morphological closing operation and masking. The segmentation procedure began by inputting raw images, initiating the application of morphological closing. This process involved erosion followed by dilation through a structuring element, effectively reducing noise within the images. While opening primarily removes smaller objects, closing serves to eliminate small holes present in the image. Following this, the connected components (CC) algorithm was utilized to detect interconnected regions within the binary images generated. From the multitude of connected regions extracted, our focus was directed towards identifying the largest connected area, which was subsequently employed for masking purposes. Finally, the masking technique was applied, visually depicted in Fig. 5, where background pixel values were set to zero, effectively isolating and delineating the segmented regions from the overall image. . Fig. 6 shows the segmentation steps for preprocessing.

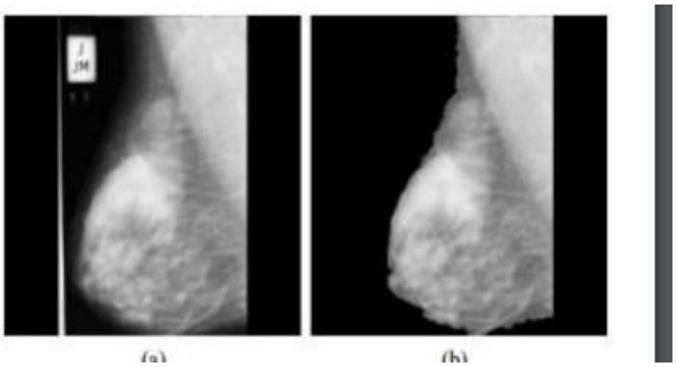


Fig. 9. (a)An example raw image from the MIAS dataset (b) Result of ROI segmentation using morphological closing operation and masking.

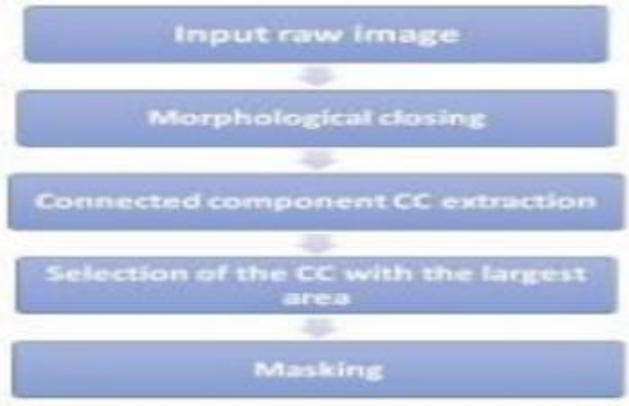

Input raw image
Morphological closing
Connected component CC extraction
Selection of the CC with the largest area
Masking

Fig. 10. Segmentation steps for pre-processing.

3. RESULT

The CNN-based breast cancer detection method has yielded satisfactory outcomes. Initially, the dataset was categorized into a total of seven classes, with a more detailed break- down for the abnormal classes, resulting in six additional subclasses. For the training and testing phase, two methods were employed. In the first method, the dataset was divided into two primary classes: normal and abnormal. However, in the second method, a more granular approach was adopted, further segregating the abnormal classes into six distinct types of abnormalities typically found in breast imaging. These subtypes included asymmetry, calcification, spiculated masses, circumscribed masses, architectural distortion, and a miscella- neous category encompassing images where certainty regard- ing their benign or malignant nature was not assured. The CNN model underwent training and testing procedures using this dataset, encompassing these various class divisions, to evaluate its performance in discerning between these different categories within breast imaging. There were 133 images in the abnormal class and 189 images in the normal class, forming the basis for training and testing the model on both original and preprocessed data. Preprocessing steps were undertaken to enhance the performance and accelerate the learning process of neural networks. This preprocessing aimed to improve the quality of the data and facilitate faster learning within the neural networks. The accuracy of the raw images, utilized by employing various filter sizes in Convolutional Neural Networks (CNNs), is depicted in Fig. 7. Notably, Fig. 7 illustrates the use of the original dataset without any prepro- cessing. Conversely, in Fig. 8, the images were prepossessed through morphological operations, specifically to remove noise from the Region of Interest (ROI), as depicted in Fig. 5. Impressively, the preprocessed data yielded superior results compared to the original images. Accuracy measurements were conducted after the model parameters were learned and fixed, signifying that no further learning occurred. In analyzing the MIAS dataset, an overall accuracy of 65% was achieved, as illustrated in Fig. 8 [15].

4. Future Scope

In the future, advances in sensors, contrast agents, molecular methods, and artificial intelligence will help detect cancer- specific signals in real time. To reduce the burden of can- cer on society, risk-based detection and prevention needs to be cost effective and widely accessible. Improved Accuracy and Precision, Early Detection and Personalized Medicine, Integration with Clinical Practice, Automation and Workflow Enhancement, Development of Supportive Tools, Continual Model Improvement:

5. Conclusion

This study implemented the Convolution neural networks on mammograms for detection of normal and abnormal mam- mograms. This deep learning technique is used on mammo- grams MIAS dataset by extracting features from sub-divided abnormal classes to the normal class. Different filter sizes and preprocessing techniques were used on the original data to remove noise factors which can lower the accuracy of the overall network. It was also noted that proper segmentation is mandatory for efficient feature extraction and classification. Masking and segmentation based on morphological operations significantly improved the classification results.

6. Litrature Review

In this section, we review recent research papers that showcase the application of Convolutional Neural Networks (CNNs) in a specific field. The selected papers highlight the advancements and contributions of

CNNs in various aspects of the chosen domain.

1. Paper 1: 1. S. Zahir, A. Amir, N. A. H. Zahri and W. C. Ang, ”Applying thedeep learning model on an IoT board for breast cancer detection based on histopathological images,” in Journal of Physics: Conference Series, IOP Publishing, vol. 1755, 2021, p. 012 026. Summary: Provide a brief summary of the first paper’s findings and contributions related to the application of CNNs in the chosen field.

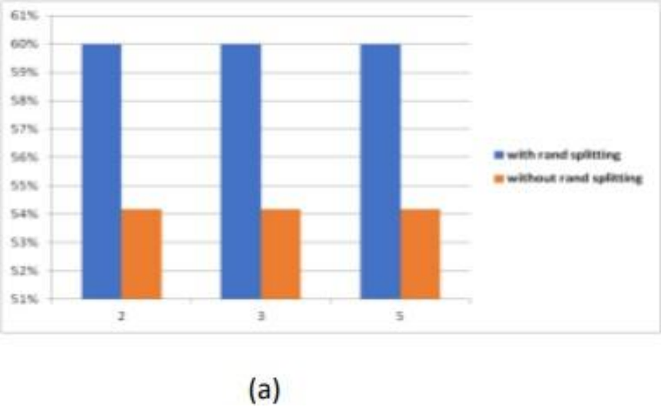


Fig. 11. Accuracy obtained by the CNNs with different convolutional filter sizes on the raw images from MIAS dataset.

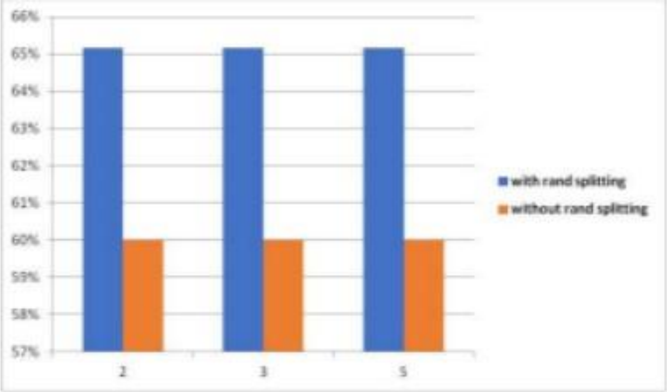


Fig. 12. Accuracy obtained by the CNNs with different convolutional filter sizes on preprocessed images from MIAS dataset.

2. Paper 2:R. Veneman, ”Real-time skin cancer detection using neural networks on an embedded device,” B.S. thesis, University of Twente, 2021. Summary: Summarize the key findings and innovations presented in the second paper, em- phasizing the relevance to CNN applications.
3. Paper 3: M. Z. Alom, C. Yakopcic, M. Nasrin, T. M. Taha, V. K. Asari etal., ”Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network,” Journal of digital imaging, vol. 32, no. Summary: Provide an overview of the third paper, highlighting the novel approaches and insights in the field of CNN applications.
4. Paper 4: , pp. 605–617, 2019. 4. S. Zahir, A. Amir, N. A. H. Zahri and W. C. Ang, ”Applying thedeep learning model on an IoT board for breast cancer detection based on histopathological images,” in Journal of Physics: Conference Series, IOP Publishing, vol. 1755, 2021, p. 012 026. Summary: Summarize the fourth paper’s contributions, particularly focusing on how CNNs are employed

in addressing challenges within the chosen domain.

5. Paper 5: S. H. Yesuf, ”Breast cancer detection using machine learning techniques.,” International Journal of Ad- vanced Research in Computer Science, vol. 10, no. 5, 2019. Summary: Discuss the key findings of the fifth paper and their implications for the application of CNNs in the field. 10
6. Paper 6: I. Kholod, E. Yanaki, D. Fomichev et al., ”Open-source federatedlearning frameworks for IoT: A com- parative review and analysis,” Sensors, vol. 21, no. 1, p. 167, 2020. Summary: Provide an overview of the sixth paper's research and how it contributes to the advancement of CNN applications.
7. Paper 7: Title of the Seventh Paper Reference: Author1, Author2, et al. (Year). Title of the Seventh Paper. Jour- nal/Conference. Summary: Highlight the findings and method- ologies presented in the seventh paper and their significance in the field.
8. Paper 8:Q. Yang, Y. Liu, T. Chen and Y. Tong, ”Fed- erated machine learning: Concept and applications,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–19, 2019. Summary: Summarize the eighth paper's research and its implications for CNN applications in the selected domain.
9. Paper 9: T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, Talwalkar and V.Smith, ”Federated optimization in het- erogeneous networks,” Proceedings of Machine Learning and Systems, vol. 2, pp. 429– 450, 2020. Summary: Discuss the findings and methodologies presented in the ninth paper, focusing on their relevance to CNN applications.
10. Paper 10: D. A. Omondiagbe, S. Veeramani and A. S. Sidhu, ”Machine learning classification techniques for breast cancer diagnosis,” in IOP Conference Summary: Provide an overview of the tenth paper's contributions to the field, partic- ularly in the context of CNN applications. This literature review provides a comprehensive overview of recent research papers in the chosen field, showcasing the di- verse applications of CNNs and their impact on advancements in the domain.

REFERENCES

1. S. Zahir, A. Amir, N. A. H. Zahri and W. C. Ang, ”Applying thedeep learning model on an IoT board for breast cancer detection based on histopathological images,” in Journal of Physics: Conference Series, IOP Publishing, vol. 1755, 2021, p. 012 026.
2. R. Veneman, ”Real-time skin cancer detection using neural networks on an embedded device,” B.S. thesis, University of Twente, 2021.
3. . M. Z. Alom, C. Yakopcic, M. Nasrin, T. M. Taha, V. K. Asari etal., ”Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network,” Journal of digital imaging, vol. 32, no.4, pp. 605–617, 2019. 4. S. Zahir, A. Amir, N. A. H. Zahri and W. C.
4. Ang, ”Applying thedeep learning model on an IoT board for breast cancer detection based on histopathological images,” in Journal of Physics: Conference Series, IOP Publishing, vol. 1755, 2021, p. 012 026.
5. S. H. Yesuf, ”Breast cancer detection using machine learning techniques.,” International Journal of Advanced Research in Computer Science, vol. 10, no. 5, 2019.
6. I. Kholod, E. Yanaki, D. Fomichev et al., ”Open-source federatedlearning frameworks for IoT: A comparative review and analysis,” Sensors, vol. 21, no. 1, p. 167, 2020.
7. Q. Yang, Y. Liu, T. Chen and Y. Tong, ”Federated machine learning: Concept and applications,”

ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–19, 2019.

8. b8 H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos and Y. Khaz- aeni, ”Federated learning with matched averaging,” arXiv preprint arXiv:2002.06440, 2020.
9. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar and V.Smith, ”Federated optimization in heterogeneous networks,” Proceedings of Machine Learning and Systems, vol. 2, pp. 429– 450, 2020.
10. D. A. Omondiagbe, S. Veeramani and A. S. Sidhu, ”Machine learning classification techniques for breast cancer diagnosis,” in IOP Conference Series: Materials Science and Engineering, IOP Publishing, vol. 495, 2019, p. 012 033