

A CNN-based Approach for Breast Cancer Classification from Ultrasound Images

Dila Marta Putri^{1*}, M Ikhsan¹, Siti Nurjanah², Fahrizal¹, Bastul Wajhi Akramunnas¹,
Asde Rahmawati¹

¹Department of Electrical Engineering, Institut Teknologi Bisnis Riau, Indonesia
Jl. Jend. Sudirman, Tengkerang Sel., Kec. Bukit Raya, Kota Pekanbaru, Riau

²Department of Electrical Engineering, Universitas Riau, Indonesia

Jl. HR Soebrantas KM. 12.5, Simpang Baru, Binawidya, Kota Pekanbaru, Riau

E-mail: putri.dilamarta@gmail.com*

Abstract

Breast cancer is one of the most frequently diagnosed cancers and remains a leading cause of cancer-related mortality among women worldwide. According to WHO Globocan 2020, breast cancer ranks second globally, with 2,262,419 cases out of a total of 19,292,289 cancer cases, accounting for approximately 11.7%. Early detection plays a critical role in reducing breast cancer mortality. In this study, a machine learning-based approach using Convolutional Neural Networks (CNN) was employed to classify breast cancer using ultrasound imaging. The dataset, collected by Al-Dhabyani et al. at Baheya Hospital in 2018, consists of ultrasound images of women aged between 25 and 75 years. The proposed CNN model includes stages of data input, preprocessing, training, testing, and performance evaluation. The model achieved an accuracy of 85%, demonstrating promising results for automated breast cancer detection. Further optimization is recommended to improve classification performance.

Keywords: Breast Cancer, Ultrasound Image, Convolutional Neural Network (CNN)

Abstrak

Kanker payudara merupakan salah satu jenis kanker yang paling sering didiagnosis dan masih menjadi penyebab utama kematian terkait kanker pada wanita di seluruh dunia. Berdasarkan data WHO Globocan 2020, kanker payudara menempati peringkat kedua secara global, dengan 2.262.419 kasus dari total 19.292.289 kasus kanker, atau sekitar 11,7%. Deteksi dini memegang peran penting dalam menurunkan angka kematian akibat kanker payudara. Dalam penelitian ini, pendekatan berbasis pembelajaran mesin menggunakan metode Convolutional Neural Network (CNN) diterapkan untuk klasifikasi kanker payudara berdasarkan citra ultrasonografi. Dataset yang digunakan dikumpulkan oleh Al-Dhabyani et al. di Rumah Sakit Baheya pada tahun 2018, yang terdiri atas citra ultrasonografi payudara dari wanita berusia antara 25 hingga 75 tahun. Model CNN yang dikembangkan mencakup tahapan input data, praproses, pelatihan, pengujian, dan evaluasi kinerja. Model ini mencapai akurasi sebesar 85%, yang menunjukkan hasil yang menjanjikan untuk deteksi otomatis kanker payudara. Optimalisasi lebih lanjut disarankan untuk meningkatkan kinerja klasifikasi..

Kata kunci: Kanker Payudara, Citra Ultrasound, Convolutional Neural Network (CNN)

1. Introduction

Based on the 2020 data from the International Agency for Research on Cancer (IARC), WHO Globocan, breast cancer ranks as the second most common cancer worldwide, with 2,262,419 cases out of a total of 19,292,289, accounting for approximately 11.7% of all cancer cases globally. In Asia, breast cancer exhibits the highest cancer-related mortality rate, reaching 50.5%. These statistics underscore the critical need for more effective, timely, and

comprehensive approaches to breast cancer detection, treatment, and management.

Malignant-stage breast cancer requires prompt and effective treatment due to its severe impact, which can lead to patient mortality [2]. A significant proportion of breast cancer-related deaths occur because approximately 70% of patients are diagnosed at an advanced stage. However, it is estimated that around 43% of cancer-related deaths could be prevented through regular early detection and avoidance of known

cancer risk factors [7]. Therefore, early detection serves as a secondary prevention effort for breast cancer, aiming to reduce mortality rates associated with the disease [6]. Ultrasonography (USG) can be utilized as an initial assessment tool for organ systems, as ultrasonic waves are effective in distinguishing various tissue structures without involving ionizing radiation. Additionally, USG is more cost-effective compared to other imaging modalities such as MRI. Due to its safety, simplicity, non-invasive nature, and widespread availability in hospitals, USG is considered a suitable imaging technique for early breast cancer detection [3].

Machine learning has been widely utilized to assist medical experts in predicting breast cancer. Various methods, such as K-Nearest Neighbor (KNN) and Naive Bayes, have been applied in previous studies to classify the Wisconsin Breast Cancer dataset. The study compared the classification accuracy of both methods, showing that the KNN method achieved an accuracy of 96.2% using K-fold cross-validation, while the Naive Bayes method reached a higher accuracy of 97.5% [8].

A similar study on breast cancer detection using ultrasound imaging employed the Convolutional Neural Network (CNN) method. In a study conducted by Liu et al. (2018), the prediction accuracy using the baseline model reached 82.9%. Meanwhile, the multi-task learning approach showed a slight improvement in prediction accuracy, achieving 83.3%.

The study referenced in this paper utilizes an open-access dataset obtained from the Kaggle platform. The dataset was originally compiled at Baheya Hospital for Early Detection and Treatment of Women's Cancer in Cairo, Egypt, in 2018, and consists of ultrasound images accompanied by verified cancer diagnoses. The dataset was processed using a CNN-based approach to classify the malignancy level of breast cancer into three categories: normal, benign, and malignant, with the aim of improving classification accuracy.

Unlike prior studies, the novelty of this research lies in the integration of data augmentation, image resizing, and label-specific data balancing techniques, along with the deployment of a simplified yet optimized CNN architecture. This architecture is specifically designed to extract subtle and discriminative features from ultrasound images, which often pose a challenge due to noise and low contrast. In addition, a lightweight deployment feature was implemented to enable real-time prediction via a user interface. The primary objective of this study is to enhance the accuracy and usability of

breast cancer classification systems by developing a robust and efficient CNN model tailored for ultrasound imaging data, thereby providing a supportive tool for early clinical diagnosis.

2. Methodology

This chapter describes in detail the methodology employed in this study to detect breast cancer based on its malignancy classification. The detection is performed using ultrasound (USG) images processed through a deep learning approach, specifically the Convolutional Neural Network (CNN) method.

2.1. Data Set

The dataset utilized in this study was collected by Walid Al-Dhabyani, Mohammed Goma, Hussien Khaled, and Aly Fahmy. This dataset comprises breast ultrasound images from women aged 25 to 75 years. The data collection took place in 2018 at Baheya Hospital for Early Detection & Treatment of Women's Cancer, Cairo, Egypt. A total of 600 female patients contributed to the dataset, which consists of 780 images. The images have an average resolution of 500×500 pixels and are stored in PNG format. Ground truth images are provided alongside the original images. The images are categorized into three classes: normal, benign, and malignant. Specifically, the dataset includes 133 normal ultrasound images, 437 images of benign cancer, and 210 images of malignant cancer. An example of the breast cancer ultrasound image dataset is shown in Figure 1.



Figure 1. Breast Cancer Ultrasound Image Dataset [1].

2.2. CNN Classification

The method employed in this study is Convolutional Neural Network (CNN), which is used to classify breast ultrasound images into three categories: benign, malignant, and normal. Several libraries are utilized during the process, including NumPy, Pandas, and Seaborn for data visualization, and OpenCV (cv2) for reading ultrasound images. The implementation uses TensorFlow, incorporating the Image Data Generator to augment the dataset in order to increase data variability and reduce the risk of overfitting prior to training the CNN model. Data shuffling is also applied to randomize the input.

In addition to the aforementioned libraries, various other libraries were used to support the development and execution of the program.

The next step is color declaration, which is carried out to avoid difficulties in calling the hex codes for colors. Instead, colors can simply be referenced using array indices. In this program, the dataset was labeled to indicate the severity level of breast cancer: benign, malignant, and normal. These labels were used when splitting the data into training and testing sets, as well as for predicting the severity level of the cancer based on the obtained results. The data was divided into training and testing sets according to each label. Additionally, image sizes were resized to 150×150 pixels to reduce computational time. Subsequently, the data was loaded based on each sample or label and shuffled to ensure that the training and testing sets were randomized during every execution. The process then continued with the model development by dividing the data into training and testing datasets.

In the final stage of the program implementation, the author added a widget button that allows users to upload breast ultrasound images. After uploading, the user can click the Predict button to classify the uploaded image using the trained CNN model.

2.3. System Design

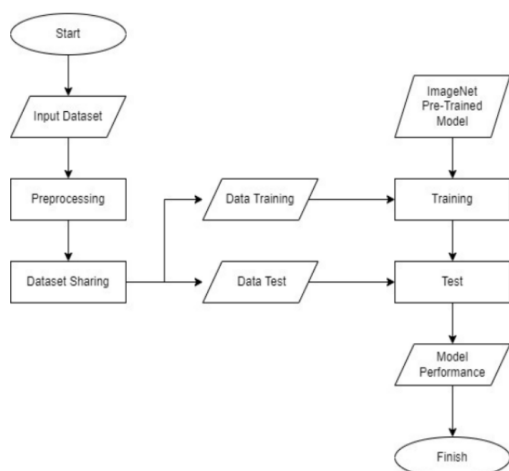


Figure 2. Flowchart System.

Figure 2 illustrates the system's workflow, detailing each stage. The first stage involves data input. Here, the existing dataset is uploaded to Google Drive, which is then connected to the program running on Google Colab. The second stage is pre-processing, which includes data validation and imputation. Validation aims to assess the completeness and accuracy of the filtered data.

The third stage is dataset splitting, where the dataset is divided into two groups: training and testing. The training data comprises 90% of the total dataset, with the remaining 10% allocated for testing. After the data splitting is complete, the process proceeds to the next stage: model training. In this stage, the divided dataset is used to train a Convolutional Neural Network (CNN) algorithm. The accuracy values for each label are then visualized in a confusion matrix.

The final stage is model performance. A simple deployment of the program is implemented by creating a widget button that allows users to input and predict breast cancer images. This straightforward deployment facilitates easy testing of the breast cancer detection program, which utilizes the CNN algorithm.

3. Results and Discussion

When the data is split into two sets, a data size of 0.1, or 10% of the total dataset, is used as the testing data, while the remaining 90% is used for training. Pre-trained weights from ImageNet are imported to enable transfer learning, allowing previously learned features to be reused in the current classification task, thereby reducing processing time. The implemented model includes several layers and three dense layers representing the three classification labels. The training process is conducted over 50 epochs, followed by visualization of the processing results.

The training process was conducted using the dataset previously prepared by the author, and the author trained a Convolutional Neural Network (CNN) model. The author evaluated the model by observing the training loss, validation loss, training accuracy, and validation accuracy. The training accuracy obtained in this study was relatively high, with values approaching 1. In contrast, the validation accuracy achieved an average value above 0.8.

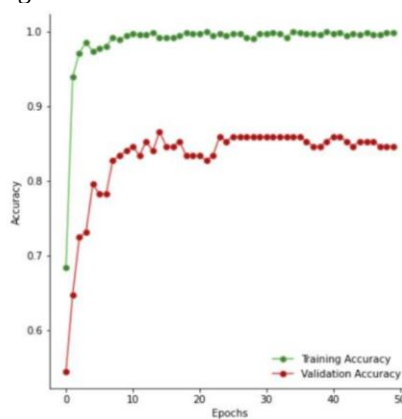


Figure 3. Training and Validation Accuracy.

Based on Figure 3, it can be observed that the training loss reached a maximum value of 0.7 and gradually decreased toward 0. Meanwhile, the validation loss reached a maximum value of 1.3 and a minimum value of 0.79, with an average of approximately 0.9.

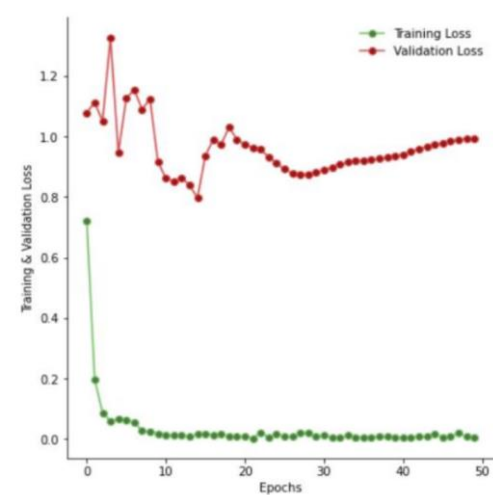


Figure 4. Training and Validation Loss.

As illustrated in Figure 4, the proposed CNN model demonstrates a cancer classification accuracy of 0.85, or 85%. This level of accuracy indicates that the model is capable of learning significant features from the dataset, enabling effective generalization to unseen data. The training loss shows a significant decrease, starting at approximately 0.7 and gradually approaching zero, indicating effective learning during training. In contrast, the validation loss fluctuates, reaching a maximum value of 1.3 and a minimum of around 0.79, with an average of approximately 0.9. The relatively large gap between training and validation loss suggests the possibility of overfitting, where the model performs well on training data but less consistently on validation data. Nevertheless, the overall performance shows that the model has strong potential in assisting cancer detection based on medical imaging, although further optimization may be necessary to improve generalization.

	precision	recall	f1-score	support
0	0.88	0.84	0.86	87
1	0.83	0.81	0.82	42
2	0.78	0.93	0.85	27
accuracy			0.85	156
macro avg	0.83	0.86	0.84	156
weighted avg	0.85	0.85	0.85	156

Figure 5. Visualization of Classification Results.

Based on the results shown in Figure 5, the next step involves grouping the precision and accuracy levels of the obtained data using the confusion matrix.

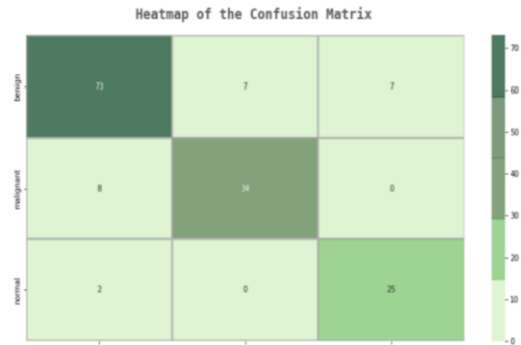


Figure 6. Heatmap of the Confusion Matrix.

The confusion matrix results shown in Figure 6 indicate that 73 images labeled as benign were correctly classified. However, 7 benign images were misclassified as malignant, and another 7 were misclassified as normal. For the malignant class, 34 images were correctly identified, while 8 malignant images were misclassified as benign, with no misclassifications into the normal category. In the normal class, 25 images were correctly classified, and 2 were misclassified as benign, with none misclassified as malignant. These results suggest that the system performs well in general, but certain misclassifications—particularly between benign and other categories—indicate the need for further refinement to improve classification precision.

After the entire program was completed, a semi-deployment phase was implemented to evaluate the results of the system. This phase utilized a button widget to allow users to upload breast cancer images and classify them according to the predefined labels.



Figure 7. Prediction Result of a Breast Cancer Image with the Benign Category.

As shown in Figure 7, the semi-deployment interface enables users to upload breast cancer images using a file upload widget. Upon clicking the "Predict" button, the system processes the input image and displays the predicted

classification result—such as "benign"—directly in the interface. This implementation allows for an intuitive user interaction and serves as a demonstration of the model's real-time prediction capability

4. Conclusion

Early diagnosis and accurate classification of breast cancer are critical in improving patient outcomes and reducing mortality rates. Prompt detection allows healthcare professionals to initiate effective treatment strategies at an early stage, increasing the chances of successful intervention. In this study, a Convolutional Neural Network (CNN) model was implemented to classify breast cancer into three categories: benign, malignant, and normal. The classification was performed using an ultrasound image-based dataset, which reflects a non-invasive and widely accessible diagnostic method. The experimental results demonstrate that the proposed CNN model achieved a classification accuracy of 85%, indicating strong potential for real-world medical applications. Although the model shows promising performance, further improvements can be made through hyperparameter tuning, increasing the dataset size, and integrating advanced preprocessing techniques to enhance overall diagnostic reliability and precision.

Daftar Pustaka

- [1] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data Brief*. 2020. Doi: 10.1016/j.dib.2019.104863.
- [2] Atthalla I N, Jovandy A, Habibie H. Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor. *Annual Research Seminar (ARS)*. 2018. Page: 148–151.
- [3] Aviana R, Astuti L, Wijaya I. Akurasi Ultrasonografi dalam Mendiagnosis Kanker Payudara di RSUP Sanglah Denpasar. *Jurnal Medika Udayana*. 2019.
- [4] International Agency for Research on Cancer (IARC). *Indonesia - Global Cancer Observatory*. Globocan. 2020.
- [5] Liu J, Li W, Zhao N, Cao K, Yin Y, Song Q, Chen H, Gong X. Integrate Domain Knowledge in Training CNN for Ultrasonography Breast Cancer Diagnosis. *Lecture Notes in Computer Science*. 2018. Page: 868–875. Doi: https://doi.org/10.1007/978-3-030-00934-2_96
- [6] Rahayu R F, Rospitasari, Prabata A, Maharina L. Karakteristik Hasil Ultrasonografi Payudara pada Program Deteksi Dini Kanker Payudara di Kota Solo. *Journal of Agri-food, Nutrition and Public Health*. 2020
- [7] Rokom. Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan dalam Sehat Negeriku KEMENKES. 2022 Feb 02. Available on <https://sehatnegeriku.kemkes.go.id/baca/umum/20220202/1639254/kanker-payudara-paling-banyak-diindonesia-kemenkes-targetkanpemerataanlayanan-kesehatan/targetkanpemerataan-layanan-kesehatan/>
- [8] Wicaksana P D. Perbandingan Algoritma K-Nearest Neighbor Dan Naive Bayes Untuk Studi Data Wisconsin Diagnosis Breast Cancer [Skripsi]. Yogyakarta: Universitas Sanata Dharma. 2015.