Convolutional neural Network-XGBoost for accuracy enhancement of breast cancer detection

E Sugiharti¹, R Arifudin¹, D T Wiyanti², A B Susilo³
¹Lecturers of Computer Science Department of UNNES
²Lecturer of Mathematics Department of UNNES
³Student of Computer Science Department of UNNES

Corresponding Author: endangsugiharti@mail.unnes.ac.id

Abstract. Computer programs can work by imitating the human brain to make decisions that can be used in the health sector. One of them is the Convolutional Neural Network (CNN) which is combined with XGBoost as the classifier. CNN-XGBoost can be implemented for the accuracy of early detection of breast cancer. The problem is how to improve the accuracy of breast cancer detection on mammogram images. The stages of the research method: (1) Collecting the MIAS 2012 dataset, (2) Dividing data into training data and testing data. (3) preprocessing: cropping, resizing, and reshaping. (4) Data Augmentation (5) Transfer Learning (6) Classification using CNN-XGBoost (7) Testing the accuracy. Based on the research that has been carried out, the results are: (1) Obtained data on the accuracy of using CNN-XGBoost on mammogram image analysis in early detection of breast cancer. (2) Further testing is needed to improve accuracy. Further testing is needed with the use of other method or by improving the quality of mammogram images.

1. Introduction
The rapid development of computer technology provides more and more benefits to human life. Now, computers can work by imitating the human brain to make decisions that can be used in the health sector. One of its uses is the detection of breast cancer. Nowadays data and information on tumor features are increasing dramatically, so a number of researchers are turning to data mining technology and machine learning approaches to predict breast cancer.

Machine learning is defined as an experience-based computational method for improving performance or making precise predictions. The experience here is previous information that has been available and can be used as student data, [1]. Technically machine learning looks for an important representation of some input data, in a predetermined probability space, using the guidance of the feedback signal. From these ideas it is possible to complete a wide range of intellectual tasks, from speech recognition to driving a car automatically, [2].

One of the machine learning techniques used is deep learning in which computer models perform classification tasks directly learning from text, images, or sound. The essence of deep learning is to automate the process of discovering of features or efficient machine learning representations for different tasks, including the simultaneous transfer of information automatically from one task to another [3]. A machine learns to grade directly from pictures, text, or sounds in deep learning. When a computer is trained to use many labeled datasets, an image’s pixel values are then transformed into an internal representation or feature vector in which classifiers can detect or identify input patterns. A large
number of datasets and a Convolutional Neural Network (CNN) architecture comprising multiple layers are trained on models. In medical imaging, deep learning is used to automatically classify cancer cells [5].

A type of deep learning that is widely used in image data is the Convolutional Neural Network (CNN). CNN may be used to detect objects in an image and recognize them. CNN is not that different, generally speaking, from the normal neural network. CNN is made up of neurons with weight, bias, and activation functions [6]. CNN is one of the most common deep learning algorithms, a kind of machine learning, in which a model learns directly from image, video, text, or voice to perform classification tasks. For finding patterns in images to identify objects, faces, and scenes, CNN is very useful. CNN learns from image data directly, uses patterns to identify images, and removes the need for manual extraction of features.

The layers of the neural network have a 3-dimensional arrangement of neurons (width, height, depth). Like other neural networks, CNN consists of an input layer, an output layer, and a hidden layer. Look at figure 1 below.

![Figure 1. The neural network layer](image)

CNN may have tens to hundreds of layers, each learning to detect different pictures. Image processing is applied at a different resolution to each training image, and the output of each image is processed and used as input to the next layer as shown in Figure 1. Image processing may begin with very simple characteristics, such as brightness and edges or increase the complexity of characteristics that identify objects uniquely according to layer thickness [7].

Data Augmentation is the process of generating samples by changing the training data, with the aim of increasing the accuracy and robustness of classifiers [8]. Data augmentation is equivalent to fantasy or dreaming. Humans envision different experience-based situations. Imagination enables us to obtain a better knowledge of our world [9]. Augmentation can increase the accuracy of the trained CNN model because the augmentation model gets additional data that can be useful for making models that can generalize better.

In terms of the lack of a dataset, transfer learning can be used to overcome it, as [10], who in his research revealed that one of the other alternatives to overcome the inadequacy of the dataset is to implement transfer learning. Transfer learning is a process or technique that uses a model trained on a dataset to solve other similar issues by using it as a starting point, changing and updating its parameters to match the new dataset [11].

There are now many pre-trained models available. The resulting model will learn basic classifications such as recognizing shapes, edges, etc. From pre-trained models and many more complex classifications such as color, height, classes that can be learned from newly added layers. The last few layers of the pre-trained layer must be removed or the model will learn to classify things previously trained, [12]. Transfer Learning can provide solutions to many deep learning problems when the available dataset is scarce.

According to [13] eXtreme Gradient Boosting (XGBoost) is an algorithm for regression and classification that applies the weak predictor principle of the ensemble and uses a decision tree in general. By optimizing the loss function value, optimization is performed using the boosting process. A model assessment mechanism is the loss function.
In several fields, XGBoost has been commonly used to achieve state-of-the-art results on multiple data challenges, which is a scalable and efficient machine learning method for tree boosting, [14]. The Gradient Booster algorithm, which has the advantages of fast speed and high accuracy is a fast implementation of XGBoost. This XGBoost Classifier is applied to CNN top level and will have image classification outcomes.

In their research, [15] found that the CNN-XGBoost combination gave the highest accuracy results compared to using only CNN. Through the implementation of CNN on the Mammogram Image, it is hoped that it can be used to examine the accuracy of breast cancer detection.

Based on the above background, the problem formulation is as follows. How to improve the accuracy of breast cancer detection on mammogram images? The goal to be achieved is to reveal the stages of CNN-XGBoost implementation on mammogram images through the use of augmentation and transfer learning data in testing the accuracy of breast cancer detection.

2. Methods
2.1 Subjects of Study
The subjects of this study used the MIAS 2012 dataset, which was taken from the Mammographic Image Analysis Society's Digital Mammogram Database website. The data used are mammography images with ppm format totaling 322 image data and divided into 7 classes, namely normal, and 6 abnormal types found in the breast, namely Asymmetry, Calcification, Spiculated masses, Circumscribed masses, Architectural distortion and Miscellaneous.

2.2 Research Stages based on R&D
Methods or ways to achieve the goals set using R & D, with 9 stages of checking/validation, [16]. The R&D stages used are as follows. (1) Identify the data that needs to be collected. (2) Arrangement of instruments so that the data obtained is accurate and valid. (3) Dataset Collection. (4) Focus Group Discussion 1 to test the accuracy of the instruments and analysis of the datasets that have been obtained (5) Analysis of system requirements. (6) Model-making in Google Colab starts from splitting data into training data and testing data, data preprocessing, data augmentation, transfer learning, training and validation processes and taking the last layer of training to be implemented in the XGBoost Classifier. (7) Making a breast cancer detection application system that implements a model that has been trained on Google Colab so that it can work to detect breast cancer on inputted data (8) Testing the system (9) System improvement based on trial results.

3. Results and Discussions
3.1 Results
The disclosure of the implementation stages of CNN-XGBoost on mammogram images through the use of augmentation and transfer learning data for breast cancer detection is as follows. the first stage is system requirements analysis. Here, a plan for the work stages of the system is obtained. It is then followed by data collection. The data used in this study is the Mammographic Image Analysis Society (MIAS) 2012 which consists of 7 classes, namely Architectural distortion totaling 19 samples, Asymmetry class totaling 15 samples, Calcification class totaling 25 samples, Circumscribed masses class totaling 23 samples, Miscellaneous class totaling 14 samples, Normal class totaling 206 samples, and Spiculated masses class totaling 20 samples. Then dividing data into training data and test data with composition of 70% training data and 30% testing data.

The next process is modeling at Google Colab, namely by connecting Google Colab to Google Drive to access the dataset that has been previously uploaded into Google Drive. After that the preprocessing stage is done, namely cropping, resizing, and reshaping. Cropping in this dataset aims to give more focus to the image area needed for the training stage, resize to change the image size and reshape is that each
mammographic image will be converted into a matrix containing numbers then the matrix dimensions will be modified according to what the model needs.

It then continued with data augmentation stage which uses ImageDataGenerator owned by Keras library to augment data where the number of images generated by ImageDataGenerator follows the number of epochs. The result of data augmentation can be seen in figure 2. After that transfer learning is done which used as a feature extractor. This study uses an architectural model ResNet50 which has been trained in ImageNet data by creating an architecture that is identical to ResNet50 but without the full connected layer and the download weight.

![Before augmentation](image1.png) ![After augmentation](image2.png)

**Figure 2.** Image before and after augmentation process

Training and validation are also done. After the model is successfully trained and gets the results of training and validation accuracy, the model trained with CNN is analyzed using confusion matrix, and the last layer of the model is taken to be implemented into the XGBoost classifier. After the fitting process, then the model can generate a training score and validation score with the XGBoost Classifier. The results of the training score for 7 classes produce a maximum value of 100%, while for 2 classes only 61% and validation scores for 7 classes with a value of 58%, and for 2 classes 74%. This result can be seen in Table 1 as follows.

| Table 1. Result of loss and accuracy value |
|-------------------------------------------|
| Training data Accuracy | Validation data Accuracy |
|-------------------------|--------------------------|
| CNN+ResNet50 7 classes  | 0,71                     | 0,51                     |
| CNN+ResNet50+XGB 7 classes | 1,00                  | 0,58                     |
| CNN+ResNet50 2 classes  | 0,61                     | 0,74                     |
| CNN+ResNet50+XGB 2 classes | 0,61                  | 0,74                     |

The final stage is system implementation. At this stage, the application is made for testing the methods carried out in this study. Web-based application. This application is made using HTML, CSS and Javascript programming language to create a user interface design. Meanwhile, for processing using the Python programming language with the Django framework. At this stage the system can detect breast cancer in the input image.

3.2 Discussions

The conclusion of this study is about the stages of CNN-XGBoost implementation on mammogram images through the use of data augmentation and transfer learning in enhancing the accuracy of breast cancer detection.

Kotsiantis used traditional CNN architecture to detect breast cancer with a total of 7 classes [1]. Initially, the dataset was divided into 2 classes, namely normal and abnormal, then dividing the abnormal class into 6 classes. with the MIAS 2012 dataset. Getting an average accuracy rate of 65%. While [17] succeeded to increase the accuracy by 18.3% by implementing transfer learning using the AlexNet, GoogleNet, and ResNet architectural models. Furthermore [15] used CNN as feature extraction and
XGBoost as classifier to improve the accuracy of the MNIST and CIFAR-10 databases with an accuracy of 99% for MNIST and 80% for CIFAR-10.

Thus, the detection of breast cancer through CNN-XGBoost on Mammogram Images can be used to improve its accuracy as resulted in this research. From the results of training and validation, it can be seen that the results of validation accuracy for 2 classes are higher than the results of validation accuracy for 7 classes, but in 7 class classifications the increase in accuracy of the ResNet50 Model using XGBoost Classifier is greater than the 2 class classification. The increase in accuracy on the ResNet50 model with XGBoost Classifier for 7 classes is 7%, while for 2 classes it is 0%. The result is also affected by the quality and the number of images used. On the other hand, with the current technological advances, any technological findings will be able to be improved by the next technological findings. Likewise, the detection of breast cancer through CNN-XGBoost on Mammogram Images also allows the accuracy to be improved.

4. Conclusions

Based on the results and discussion as described above, the following conclusions can be drawn. The accuracy of breast cancer detection on mammogram images can be improved by implementing CNN-XGBoost through the use of data augmentation and transfer learning. The stages of CNN-XGBoost implementation on mammogram images begins with dataset collection. Followed by model making in Google Colab starts from splitting data into training data and testing data, data preprocessing, data augmentation, transfer learning, training and validation processes and taking the last layer of training to be implemented in the XGBoost Classifier. Finally a breast cancer detection application system was made which implements the model that has been trained on Google Colab so that it can work to detect breast cancer on inputted data.

References

[1] Kotsiantis S B, Zahar I, and Pintelas P 2007 Emerg. artif. intellig. appl. comput. eng. 31 249
[2] Fails J and Olsen J 2003 Proc. 8th int. conf. Intell. user interfaces 1 39
[3] Deng L and Yu D 2014 Signal Process 7(4) 197-387.
[4] LeCun Y, Bengio Y, and Hint G 2015 Nat 521(7553) 436-444.
[5] Charan S, Khan M J, and Khurshid K 2018 Int. Conf. Comput. Math. Eng. Technol. (iCoMET) 1 1
[6] Dong C, Loy C C, and Tang X 2016 Eur. conf. comput. vis. 9906 391
[7] Sainath N, Mohamed A, Kingsbury B, Ramabhadran B 2013 IEEE int. conf. acoust. speech signal process. 1 8614
[8] Frid-Adar M, Klang E, Amitai M, Goldberger J, and Greenspan H 2018 IEEE 15th int. symp. biomed. imaging (ISBI 2018) 1 289
[9] Salamon J and Bello J P 2017 IEEE Signal Process. Lett. 11 279
[10] Barman R, Deshpande S, Agarwal S, and Inamdar U 2019 Proc. Natl. Conf. Mach. Learn., 26th March 2019. ISBN: 978-93-5351-521-8
[11] Long M, Zhu H, Wang J, and Jordan M I 2017 Proc. 34th Int. Conf. Mach. Learn. 70 2208
[12] Shin H, Roth R, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, and Summers R 2016 IEEE trans. med. Imaging 35(5)1285
[13] Chen T and Guestrin C 2016 Proc. 22nd acm sigkdd int. conf. knowl. discov. data min.. 8 785
[14] Zheng H, Yuan J, and Chen L 2017. Energies. 10(8) 1
[15] Ren X, Guo H, Li S, Wang S & Li J 2017 Int. Workshop Digit. Watermarking (pp. 378-390). Springer, Cham.
[16] Sugiyono 2016 Metode Penelit. Kuantitatif Kualitatif dan R D. Bdg.: Penerbit Alf.
[17] Ahmad H M, Ghuffar S, and Khurshid K. 2019 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST) IEEE (pp. 328-332).