

Exploiting Silhouette Principle Component For Dimension Reduction In Breast Ultrasound Images Classification

Etika Kartikadarma ^{a,1}, Ahmad Zainul Fanani ^{a,2}, Pujiono ^{a,3}, Affandy ^{a,3}, Sari Ayu Wulandari ^{b,3}

^a Fakultas Ilmu Komputer Universitas Dian Nuswantoro Semarang, Indonesia

^b Fakultas Teknik Universitas Dian Nuswantoro Semarang, Indonesia

¹etika.kartikadarma@dsn.dinus.ac.id, ²a.zainul.fanani@dsn.dinus.ac.id, ³pujiono@dsn.dinus.ac.id, ⁴affandy@dsn.dinus.ac.id,

⁵sari.wulandari@dsn.dinus.ac.id

* corresponding author

ARTICLE INFO

Article history

Received

Revised

Accepted

Keywords

PCA,

Reduction ,

Classification

ABSTRACT

This paper proposes the use of the Dimensional Reduction method with the Silhouette Principle Component (SPC) Approach to improve the classification of breast ultrasound images. The PCA method is used to reduce the dimensions of medical images to improve classification, with MobileNet-v2 and DenseNet-121 as the optimal classification algorithm choices. The results show that the SPC method succeeded in producing efficient feature representation with data sizes that are almost the same as the original data, while PCA produces greater dimensionality reduction. The SPC model also shows the best performance in terms of accuracy and loss. This research makes a significant contribution to the development of more sophisticated and efficient medical image analysis techniques to support the diagnosis and treatment of breast cancer..

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. Introduction

Deep learning, as one of the currently trending methods in the field of artificial intelligence, highlights the complexity of the process which requires not only dimensionality reduction but also large datasets as its main basis. Dimensionality reduction is a critical aspect not only in configuring effective deep learning architectures but also an important requirement in dealing with health images that often have high dimensions and high resolution. In this context, health images have become the focus of increasing research in the application of deep learning techniques for diagnosis, treatment and medical research. This phenomenon indicates a significant paradigm shift in the approach to health, offering great potential for breakthroughs and innovations that could change the landscape of treatment and healthcare in the future.

The operation of an image resizer in some software generally involves manipulation of the image size or resolution, which can be done through various methods such as interpolation, cropping, or scaling. The interpolation method is used to increase or decrease the number of pixels in an image so that its size can be adjusted. Cropping cuts an image to keep only certain desired parts. Scaling involves changing the scale of an image proportionally. Although image resizers are often used in many cases to adjust image size or resolution, they are less optimal for images in a classification context because this process can remove or obscure important details in the image. In classification, machine learning or deep learning models tend to rely on features or patterns in images to differentiate between different classes or labels. If the image resolution is too low or the size is too small, important details can be lost or become difficult to recognize, which in turn can affect the model's ability to perform classifications accurately.

Several approaches that are often used by researchers are: Several approaches that are often used by researchers for the image dimension reduction process are as follows:

1. Principal Component Analysis (PCA): This method identifies the direction of maximum variance in the data and reduces dimensionality by projecting the data into a lower dimensional space consisting of the principal components that contribute most to the data variance.
2. Linear Discriminant Analysis (LDA): LDA tries to find a linear combination of features that maximizes the distance between classes while minimizing the variance within each class, thereby producing optimal features for classification.
3. Independent Component Analysis (ICA): ICA tries to separate a mixed signal into independent components that express the original data, thereby, reducing the image dimensions while preserving important information.
4. Autoencoder: An autoencoder is an artificial neural network that attempts to reconstruct its input in its output, but with a narrower structure in the hidden layers. This results in a denser representation of the data, which creates a dimensionality reduction effect.
5. Non-negative Matrix Factorization (NMF): This method aims to separate a non-negative matrix into two matrices with a multiplication result that is close to the original matrix. NMF has been used in image processing for feature extraction and dimension reduction.
6. Wavelet Transform: This technique breaks down the image into several scales and directions, producing coefficients that describe important features in the image. Less significant coefficients can be ignored or compressed, which reduces the image dimensions.
7. Local Binary Patterns (LBP): LBP is a method for extracting texture features from images by comparing pixel values with their neighbors. Features extracted from LBP can be used for simpler, low-dimensional image representation.

All of these methods have their respective advantages and disadvantages depending on the application context and the nature of the data being dealt with. However, PCA is often one of the most widely used methods for several reasons, namely it is simple and easy to understand, effective for dealing with variable relationships: PCA, fast processing and wide use. Although the PCA algorithm for dimensionality reduction provides excellent results, the results are still not satisfactory. In fact, annotation performance is limited in two main aspects:

1. Dimensionality reduction is influenced by manual selection of n-components so that dimensionality reduction results are often difficult to interpret intuitively.
2. Not all images in one dataset class have the same number of components (n-components). This causes additional challenges in using PCA for dimensionality reduction, because PCA assumes that each sample has a similar structure in the feature space. Variability in the number of components between images can cause important information to be lost in the dimensionality reduction process.

Therefore, to handle this kind of case, it is necessary to consider other more flexible approaches or adapt PCA with other methods to overcome the non-uniformity of the number of components in the dataset. The main objective of this paper is to develop an effective PCA model for dimensionality reduction as well as improve the performance of medical image classification by utilizing this approach. The contributions of this paper are threefold:

1. The impact of the measures studied:
Initial size variations: 256 and 512 [1]
N-component variations: 3, 5 and 10. (itself based on experiment)
For the dimension-classifier, the influence of different image sizes is analyzed;
2. Impact of decreasing input image resolution on all 4 classifiers studied (Inception-v2, DenseNet-121, ResNet-50, MobileNet-v2) [2,3]
3. Image size and PCA feature extraction approaches by selecting a certain number of principle components, as image dimensionality reducers to classify images, are proposed. This model is proven to outperform regular PCA classifiers.

The remainder of this paper is organized as follows. Section 2 introduces the motivation of related previous work reduction PCA image . Section 3 explains the details of our proposed SPCA algorithm framework . Section 4 compares our framework with existing models and analyzes the results. Section 5 concludes this paper.

2. Method

2.1 PCA Image Reduction

Dimensionality reduction in images is needed in image classification operations using deep learning, because the large number of features or attributes in medical images can cause complex and heavy computational problems. One approach to overcome this is to use feature extraction. The linear feature extraction method most often used in dimension reduction systems is Principle Component Analysis (PCA). PCA transforms data into lower dimensions, which are spatial dimensions. The linear method assumes that the image is in a lower dimensional linear space or called subspace [3,4]. PCA is one of the oldest and most commonly used linear feature extraction methods, which is widely known as a dimensionality reduction technique [5]. The concept behind this method is to reduce the dimensions of the principal components, selecting components with a large standard deviation to capture as much information as possible [6]. PCA has greater advantages compared to other dimensionality reduction methods, such as t-SNE (t-Distributed Stochastic Neighbor Embedding) and LDA (Latent Dirichlet Allocation) [7]. PCA has several advantages that make it effective in dimensionality reduction:

1. Reducing the number of features in a dataset without losing important information
2. Building a new coordinate system that accounts for the largest data variance
3. Has an easy and efficient algorithm
4. Able to reduce dimensions with the aim of ensuring a balance between information and complexity

PCA has been used in various applications, such as data processing in the field of statistics [8], data analysis in the banking field [9], and data processing [7,10]. PCA was introduced with the idea that each principal component represents a linear combination of input variables, generated by a linear function that maximizes differences and is uncorrelated with the other components. The disadvantage of PCA is that dimensionality reduction is only carried out based on data variations, so that information that is relevant but not related to data variations may be lost or be called the sum of n-components. This becomes a problem when each image in the same dataset has a different number of n-components from each other. Additional methods are needed to determine the appropriate number of n-components for each image, one of which is by measuring the level of similarity using the silhouette component/SC method [11], hereinafter referred to as Silhouette Principle Component (SPC).

2.2 PCA for dimensionality reduction

Linear feature extraction methods such as Principal Component Analysis (PCA) form data in a linear space A and transform it into a lower dimensional space. PCA is one of the oldest and most commonly used linear dimensionality reduction methods, with the idea that principal components of a smaller space can capture the most information from the data. Each principal component in PCA is a linear combination of input variables, generated by a linear function that maximizes differences and is uncorrelated with each other. The PCA stages used in this research are:

The matrix A(i, j) represents an input image in grayscale; i represents the number of rows and j represents the number of columns (Eq. 1). P is a matrix vector that represents the input image, as in equation 1.

1. Find the average value of each column A, using the equation:↵

$$\bar{A} = \sum \frac{A_i}{n} \quad \bar{B} = \sum \frac{B_i}{n} \quad \bar{C} = \sum \frac{C_i}{n} \quad \bar{D} = \sum \frac{D_i}{n} \quad (2)$$

2. Reduce each data in matrix A by its average value, so that a set of data is obtained in the reduced matrix with a new average.

$$A' = \begin{bmatrix} A_1 - \bar{A} & B_1 - \bar{B} & C_1 - \bar{C} & D_1 - \bar{D} \\ A_2 - \bar{A} & B_2 - \bar{B} & C_2 - \bar{C} & D_2 - \bar{D} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ A_{45000} - \bar{A} & B_{45000} - \bar{B} & C_{45000} - \bar{C} & D_{45000} - \bar{D} \end{bmatrix} \quad (3)$$

3. Calculating the variance covariance matrix from the equation:

$$\begin{aligned} \text{cov}(A, B) &= \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sum_{i=1}^n (A_i - \bar{A})(A_i - \bar{A})(n-1)} \\ \text{var}(A, A) &= \frac{\sum_{i=1}^n (A_i - \bar{A})(A_i - \bar{A})(n-1)}{(n-1)} \end{aligned} \quad (4)$$

4. The number of different covariance values for a dimension is . The resulting matrix from this stage is a rectangular matrix:

$$\Sigma = \begin{bmatrix} \text{cov}(A, A) & \text{cov}(A, B) & \text{cov}(A, C) & \text{cov}(A, D) \\ \text{cov}(B, A) & \text{cov}(B, B) & \text{cov}(B, C) & \text{cov}(B, D) \\ \text{cov}(C, A) & \text{cov}(C, B) & \text{cov}(C, C) & \text{cov}(C, D) \\ \text{cov}(D, A) & \text{cov}(D, B) & \text{cov}(D, C) & \text{cov}(D, D) \end{bmatrix} \quad (5)$$

5. Calculate the eigenvector and eigenvalue of the matrix using equation 6, where λ is the eigenvalue and Q is the eigenvector of the covariance matrix. Eigen values and eigen vectors can be solved using the equation:

$$\det(\lambda I - \Sigma)Q = 0 \quad (6)$$

6. With is the identity matrix. The resulting number of eigenvectors and eigenvalues will be proportional to the dimensions of the covariance matrix.
7. Calculating the feature vector obtained from the eigenvector matrix which has been sorted from the largest to the smallest eigenvector. The first column of the feature vector is the eigenvector corresponding to the first highest eigenvalue, the second column is the eigenvector corresponding to the second highest eigenvalue, and so on. In the form of a feature vector matrix, it is depicted in the form of a row matrix as follows:

$$[F] = [eigen_1 \quad eigen_2 \quad eigen_3 \quad \dots \quad eigen_n] \quad (7)$$

8. Calculate final data. The final data is obtained from multiplying the transformed feature vector with the reduced matrix with the mean using the following equation:

$$Final = [F'] \times [A] \quad (8)$$

From this final data, the principal component (PC) order will be obtained which has been sorted from low order to high order. The first column is PC1, the second column is PC2 and so on. Low order contains the most important aspects because it has the largest eigenvalue which can represent the characteristics of the data. Next, to carry out the dimension reduction process, the number of PCs or what is called the number of components (n-components) must be input to obtain dimension reduction of the image.

2.3 Propose Method

PCA still requires to input the number of components in each dimension reduction operation. In one dimension reduction, the number of n-components for each image in the dataset is equalized, even though each image has different characteristics. To find the optimal number of components for this operation, optimization techniques are used to improve accuracy. Silhouette is a way to evaluate the quality of data grouping (clustering). This is not an optimization algorithm, but rather a measure of how well the objects in a cluster resemble each other when compared to other clusters is called the SPC algorithm. This approach offers a score based on how closely the objects in one cluster match their own cluster compared to other clusters. This approach is easy to implement because

there is no need to experiment with all n-components of each image. This is more efficient and faster in determining the right n-component to use to get optimal results. After carrying out the PCA process, the results obtained are in equation 8. Next, the final phase of PCA is grouped using K-Means. After grouping is done, the next thing is done

$$A = \begin{bmatrix} A_1 & B_1 & C_1 & D_1 \\ A_2 & B_2 & C_2 & D_2 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ A_{45000} & B_{45000} & C_{45000} & D_{45000} \end{bmatrix}$$

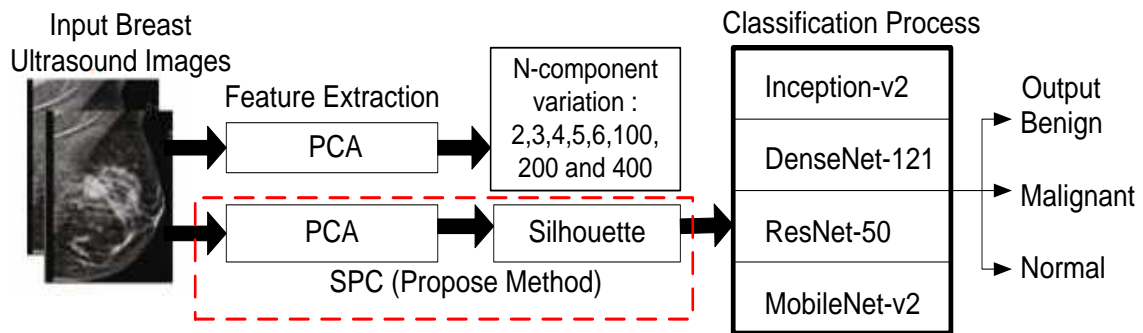


Fig. 1. The Course of Research

The profiling process results from the best grouping of pliers, namely the silhouette coefficient. An evaluation metric that is generally used in cluster analysis to determine the optimal number of clusters in data mining. This coefficient measures the extent to which an object fits its own cluster compared to its nearest cluster. The stages for calculating the silhouette coefficient are as follows:

2.4 1. Calculate the average distance of the object to all other objects in one cluster using equation 9.

$$a(i) = \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i, j) \tag{9}$$

2.5 2. Calculate the average distance of the object to all other objects in other clusters, then take the minimum value using the equation:

$$d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \tag{10}$$

3. Calculate the silhouette coefficient value with the equation:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{11}$$

Equation 11 is used to calculate the average silhouette value. The average silhouette value represents the overall clustering quality of the dataset. The silhouette coefficient value range is between -1 to 1, with a positive value indicating the degree of suitability of the object to its cluster, while a negative value indicates that the object may be more suitable in a neighboring cluster. The final result of SPC is the smallest silhouette coefficient value, meaning it is the one that is least similar to the original image but is still recognizable.

3. Results and Discussion

2.6 Experiment

The dataset we use in this research is the Breast Ultrasound Images Dataset (BUID), using a public dataset from Kaggle (<https://shorturl.at/bgkpO>). Even though the images in this dataset have moderate dimensions, the dimensions vary from one image to another. The Breast Ultrasonography Dataset dataset is categorized into three classes: normal, benign, and malignant images. The dataset collected includes ultrasound images of the breasts in women aged between 25 and 75 years. This data was collected in 2018. The number of patients was 600 female patients. The dataset consists of 780 images with an average image size of 500*500 pixels. These images are in PNG format [12]. The research scheme is shown in Figure 1.

First of all, the image dataset is subjected to a feature extraction process which consists of 2 methods, namely PCA and SPC. The feature extraction process in PCA is carried out with a variety of n-components, namely 48, 24, 12 and 6. Meanwhile in SPC, the output from PCA is forwarded to the SC calculation to get the n-component which corresponds to the maximum n-component limit of 50. The results will be compared by using the propose method, where the n-components are generated automatically with each image being different from one another. Next, the results of the extraction process were then tested using 4 classification algorithms, namely Inception-v2, DenseNet-121, ResNet-50, MobileNet-v2. The final part of the test is carried out by calculating the loss and accuracy of the classification.

3.2 Results

This research presents 2 dimension reduction algorithms which are then continued by testing the results of dimension reduction into 4 classification algorithms. The two dimension reduction algorithms are PCA and SPC (propose method). PCA is used as a comparison with the proposed method to overcome the limitations of the PCA algorithm, which relies heavily on the number of n-components. Figure 1 explains in detail what SPC suggests in this study. In fact, the functionality of PCA is comparable to the suggested approach for constructing principle component matrices. The key difference is how to generate candidate solutions, select the optimal number of n-components, and classify breast cancer types. A complete explanation of the differences between Inception-v2, DenseNet-121, ResNet-50, MobileNet-v2 is provided below.

1. Inception-v2 [2], A convolutional neural network that uses the Inception module to power feature representation with computational efficiency, enabling parallel processing with varying kernel sizes, and a batch normalization module.
2. DenseNet-121 [3], A neural network that builds connections between each layer and the next layer, produces dense feature representation and improves feature understanding.

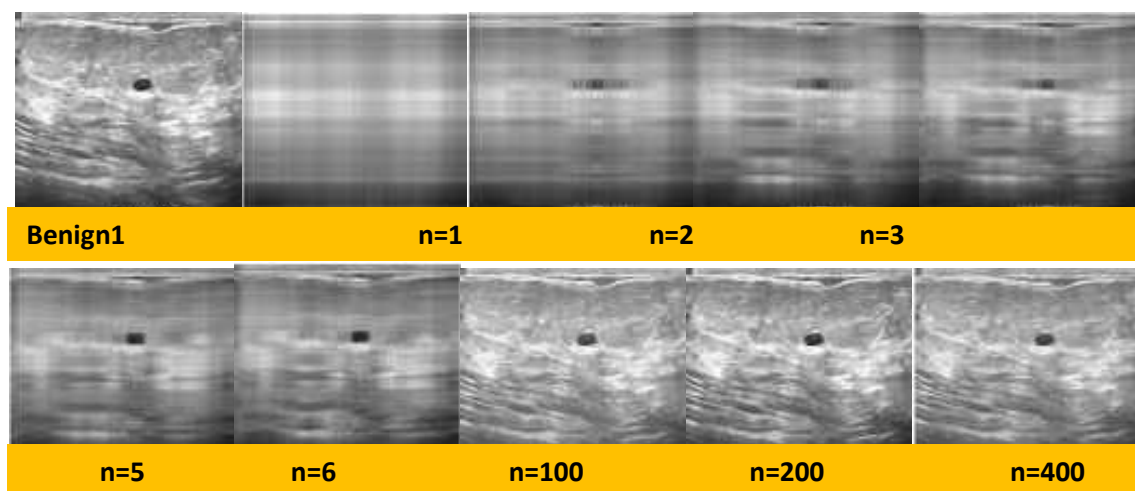


Fig. 2.Image Comparison with Different n-Components

3. ResNet-50 [2], A network architecture that uses residual blocks to address training problems in deep networks, by introducing shortcut connections to facilitate better gradient flow.

4. MobileNet-v2 [1], A lightweight network model that optimizes size and speed by using depthwise separable convolution, and introduces bottleneck layers to increase efficiency and accuracy.

The results of the qualitative comparison are shown in Figure 2. It can be seen that the classification marker for the type of breast cancer is in the black part which is located in the middle of the image. When $n=1$, the sign is not visible, at $n=2-4$, n is visible but the size becomes wider than before, the size starts to be similar to the original starting at $n=5$. So, at $n=5$, the size should be able to be used for classification. The decrease in dimensions from each change in n -component value is shown in Table 1.

Table 1. Dimensional Change

Method	Size		
	Original	Extraction	Reduce (%)
PCA-n=2	264702	942	99.64413
PCA-n=3	264702	1413	99.46619
PCA-n=4	264702	1884	99.28826
PCA-n=5	264702	2355	99.11032
PCA-n=6	264702	2826	98.93238
PCA-n=100	264702	47100	82.20641
PCA-n=200	264702	94200	64.41281
PCA-n=400	264702	188400	28.82562
SPC	264702	2355	99.11032

In this experiment, the initial data size is 264,702. By using the PCA method for feature extraction, the data size was reduced significantly, with the reduction rate increasing as the number of retained components (n) increased. However, the size of the data processed with PCA remains much larger than the size of the data processed with the SPC method. The SPC method succeeded in producing a fairly efficient feature representation with a data size that is almost the same as the original data, showing efficiency in feature extraction. The images resulting from PCA and SPC extraction for each class are shown in Figure 3.

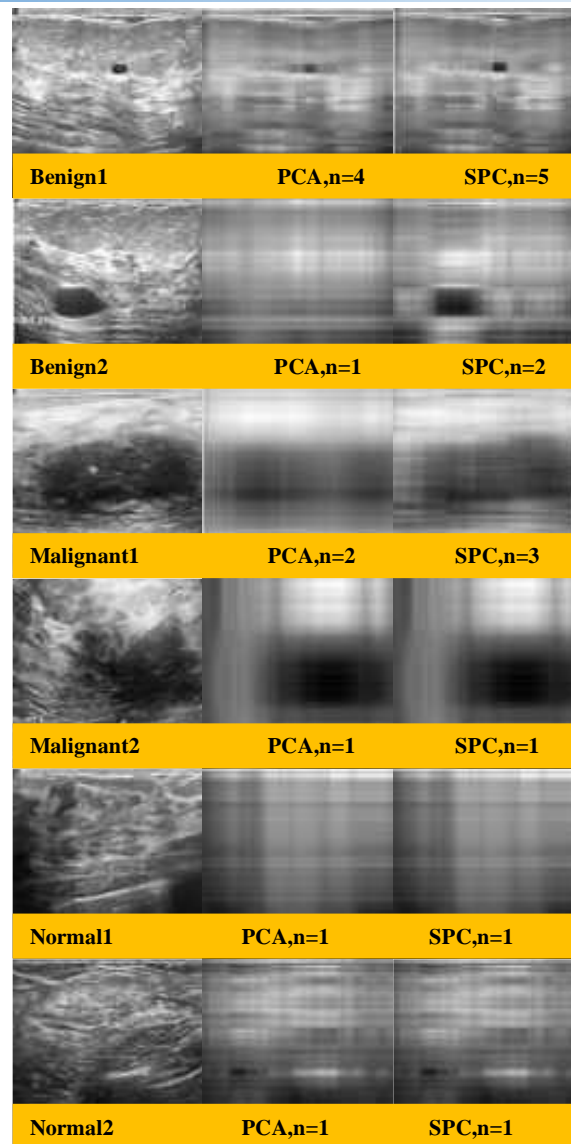


Fig. 3.Differences between PCA and SPC for each class

In Figure 3, it can be seen that the PCA before optimality still shows an image that is very different from the original image. When reaching the optimal n -component, namely in SPC, the image is similar to the original image even though it does not show similar contour details. In SPC, the optimal graph of the n -component is shown in Figure 3. This means that the silhouette score can detect the minimum size for classification, namely in the case of balign 1 this is at $n=5$.

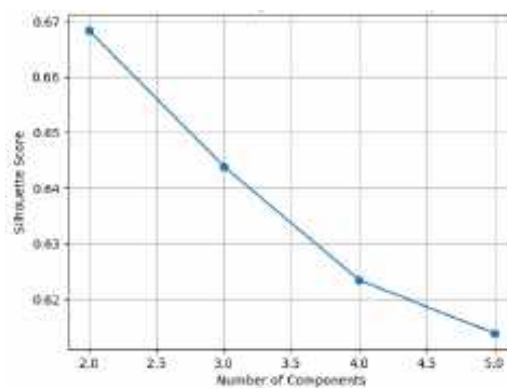


Fig. 4.Optimal Graphics from SPC

Next, the results of the SPC extraction process are used as input for 4 classification algorithms by standardizing the sizes, namely 28x28, 24x24, 12x12 and 6x6, to see the performance of the SPC compared to the original image. The loss graph for each input image dimension for each classification algorithm is shown in Figure 4.

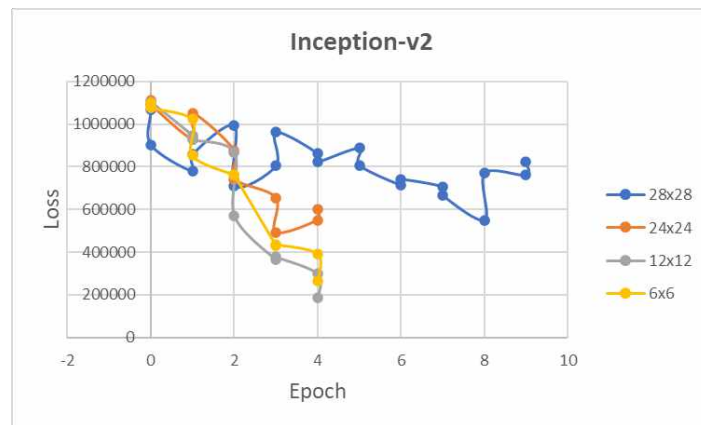


Fig. 5. Learning Data Loss Graph of Inception-v2

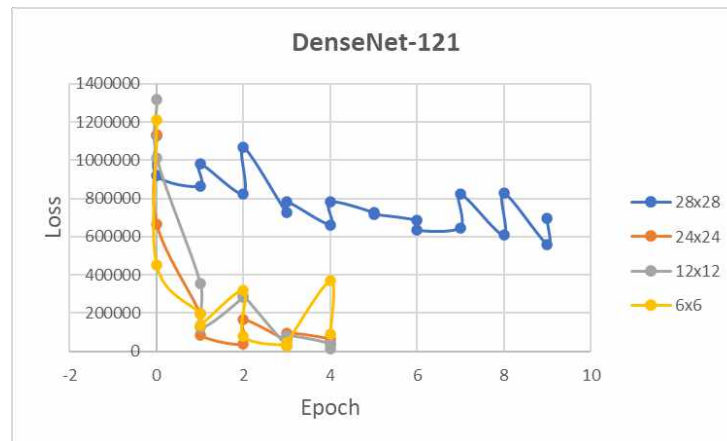


Fig. 6. Learning Data Loss Graph of (b) DenseNet-121

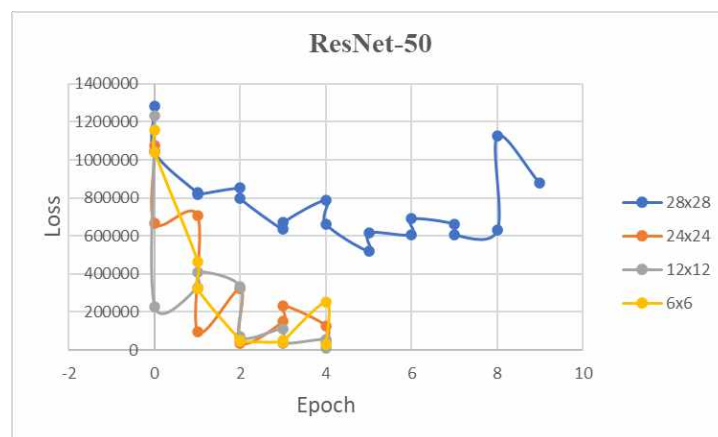


Fig. 7. Learning Data Loss Graph of (c) ResNet-50

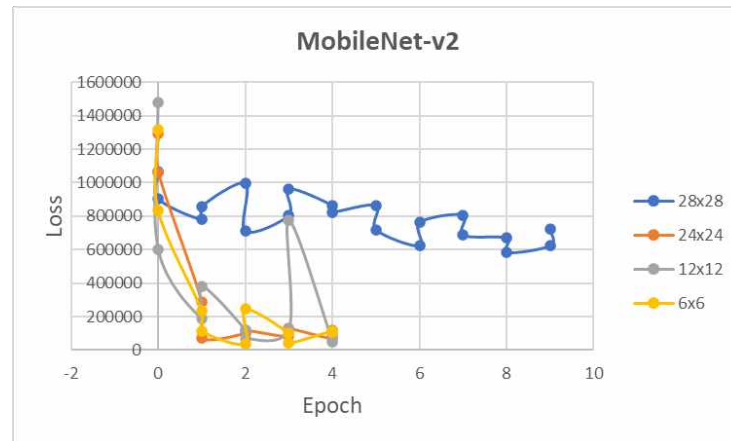


Fig. 8. Learning Data Loss Graph of (d) MobileNet-v2

The ResNet and DenseNet models show a steady decrease in loss as the number of epochs increases. Both models also achieve low loss at the end of the epoch, indicating good convergence. Meanwhile, Inception and MobileNet models show larger fluctuations in loss and tend to reach higher loss at the end of the epoch, indicating lower performance in the training process.

Table 2. Accuracy and Loss

No.	Method	Size	Accuracy	Losses
1.	Resnet-50	28X28	34%	0.0359
		24X24	98%	0.0023
		12X12	98%	0.0012
		6X6	98%	0.0034
2.	Inception-v2	28X28	77%	0.0195
		24X24	87%	0.0135
		12X12	95%	0.0058
		6X6	96%	0.0047
3.	Mobilenet-v2	28X28	50%	0.0202
		24X24	99%	0.0012
		12X12	100%	0.0003
		6X6	100%	0.0002
4.	DenseNet-121	28X28	46%	0.0195
		24X24	99%	0.0010
		12X12	99%	0.0009
		6X6	97%	0.0027

The classification test results are shown in Table 2. The MobileNet-v2 and DenseNet-121 models show the best performance in terms of accuracy and loss. MobileNet-v2 consistently achieved the highest accuracy and lowest loss across all dataset sizes tested. On the other hand, ResNet-50 and Inception-v2 show lower performance in terms of accuracy, especially on larger datasets (28x82), although Inception-v2 tends to perform better than ResNet-50. In conclusion, MobileNet-v2 and DenseNet-121 are more suitable for applications that require a balance between accuracy and computational efficiency. The worst result was on Resnet-50 at size 28x28, where the accuracy was only 34%. Figure 5 shows the confusion matrix table for the 3 classes in the training data. Some incorrect images are shown in Figure 6.

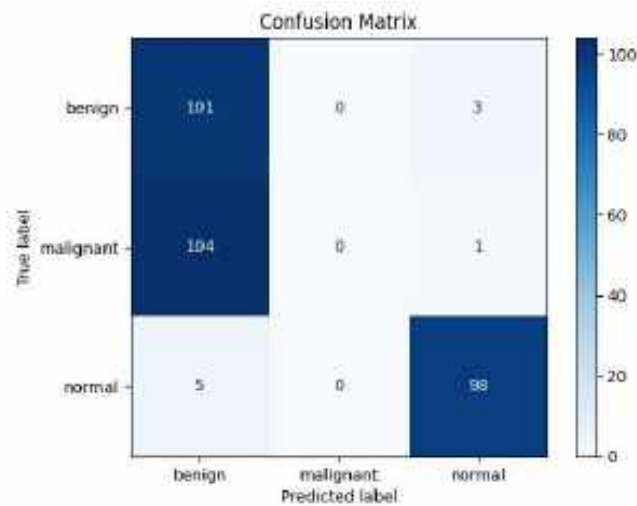


Fig. 9. Confusion Matrix from Resnet-50 with size 28x28

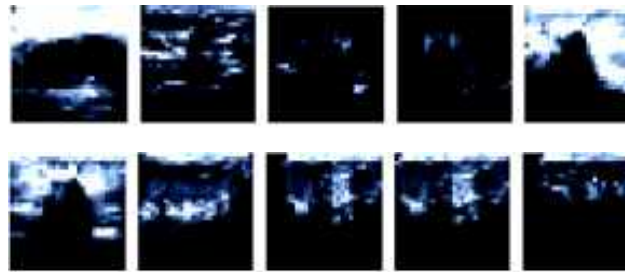


Fig. 10. Falsely Detected Image from Resnet-50 with size 28x28

4. Conclusion

This research proposes the use of the Dimensional Reduction method with the Silhouette Principle Component (SPC) Approach to improve the classification of breast ultrasound images. Through experiments carried out on the Breast Ultrasound Images Dataset (BUID) dataset using the PCA and SPC methods, and testing the results with four different classification algorithms, it was found that SPC succeeded in producing efficient feature representation with data sizes that were almost the same as the original data, while PCA produces greater dimensionality reduction. In addition, SPC automatically determines the optimal number of components for each image, thereby overcoming the limitations of PCA which relies on manual selection of the number of components. Experiments also show that MobileNet-v2 and DenseNet-121 are the best classification algorithms in terms of accuracy and loss. However, ResNet-50 and Inception-v2 also have acceptable performance especially on smaller dataset sizes. The practical implication of this research is the use of SPC as a more efficient method in reducing the dimensions of medical images to improve classification, with MobileNet-v2 and DenseNet-121 as the optimal classification algorithm choices. Thus, this research makes a significant contribution to the development of more sophisticated and efficient medical image analysis techniques to support the diagnosis and treatment of breast cancer.

References

- [1] Hossein Talebi and Peyman Milanfar. 2021. Learning to Resize Images for Computer Vision Tasks. *Computer Vision and Pattern Recognition*. <https://doi.org/10.48550/arXiv.2103.09950>
- [2] Quintana, G.I.; Li, Z.; Vancanberg, L.; Mougeot, M.; Desolneux, A.; Muller, S. Exploiting Patch Sizes and Resolutions for Multi-Scale Deep Learning in Mammogram Image Classification. *Bioengineering* 2023, 10, 534. <https://doi.org/10.3390/bioengineering10050534>
- [3] Basak, H., Kundu, R., Chakraborty, S. et al. Cervical Cytology Classification Using PCA and GWO Enhanced Deep Features Selection. *SN COMPUT. SCI.* 2, 369 (2021). <https://doi.org/10.1007/s42979-021-00741-2>
- [4] Salifu Nanga, Ahmed Tijani Bawah, Benjamin Ansah Acquaye, Mac-Issaka Billa, Francis Delali Baeta, Nii Afotey Odai, Samuel Kwaku Obeng, and Ampem Darko Nsiah. Review of dimension reduction methods. *Journal of Data Analysis and Information Processing*, 9(3):189–231, 2021.

- [5] Zhiyang Jin, Guorui Feng, Yanli Ren, and Xinpeng Zhang. Features extraction optimization of jpeg steganalysis based on residual images. *Signals processing* , 170:107455, 2020.
- [6] Xiaolu Han, Yun Liu, Zhenjiang Zhang, Xin Lü , and Which Li. Sparse auto-encoder combined with kernel for network attack detection. *Com puter Communications* , 173:14–20, 2021.
- [7] Bartenhagen, C., Klein, H.U., Ruckert, C. et al. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics* 11, 567 (2010). <https://doi.org/10.1186/1471-2105-11-567>
- [8] K. Koonsanit, D. Hiruma and N. Nishiuchi, "Dimension Reduction Method by Principal Component Analysis in the Prediction of Final User Satisfaction," 2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI), Kanazawa, Japan, 2022, pp . 649-650, doi: 10.1109/IIAIAAI55812.2022.00128.
- [9] C. Yumeng and F. Yinglan, "Research on PCA Data Dimension Reduction Algorithm Based on Entropy Weight Method," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2020, pp. 392-396, doi: 10.1109/MLBDBI51377.2020.00084.
- [10] T. Zhang and B. Yang, "Big Data Dimension Reduction Using PCA," 2016 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, 2016, pp. 152-157, doi: 10.1109/SmartCloud.2016.33.
- [11] Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy* (Basel). 2021 Jun 16;23(6):759. doi: 10.3390/e23060759. PMID: 34208552; PMCID: PMC8234541.
- [12] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data in Brief*. 2020 Feb;28:104863. DOI: 10.1016/j.dib.2019.104863