Embedding Learning for Unsupervised Breast Cancer Images Clustering

Andriamasinoro Rahajaniaina[1*], Adolphe Andriamanga Ratiarison[2]

[1]Department of Mathematics, Computer Science and Applications, University of Toamasina, Toamasina, Madagascar.

[2]Department of Physics and Applications, University of Antananarivo, Antananarivo, Madagascar.

***Corresponding Author**

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.110400082

Received: 14 April 2026; Accepted: 19 April 2026; Published: 08 May 2026

ABSTRACT

Early detection of breast cancer significantly reduces the number of deaths caused by this disease. In Africa where the number of new cases and deaths is constantly increasing. For Madagascar, very little information is available regarding the number of people affected by this disease. Advances in the application of artificial intelligence in medicine are improving the techniques for detecting this disease. Unfortunately, most of these techniques are cumbersome, complex, and very expensive. In this work, we propose a lightweight, hybrid approach to clustering breast cancer images. Our approach combines deep learning, ArcFace and unsupervised clustering. The architecture relies on the MobileNetV3Small convolutional network as a feature extractor. At the output of the backbone, a projection head is added to transform the feature maps into a compact embedding vector. The goal is to project the data into a low-dimensional (64-dimensional) latent space, where the discriminating properties between classes are strengthened. The use of ArcFace ameliorate intra-class compactness and inter-class separability, enhancing the quality of the learned representations. Two phases of training were adopted: firstly, only the projection layers and the ArcFace layer are trained, with the backbone remaining frozen to stabilize the learning process. Then, partial fine-tuning is performed by unfreezing the final layers of the convolutional neural network. Principal Component Analysis algorithm is used to facilitate the structuring of the embedding in a lower-dimensional space while preserving most of the discriminating information. A comparative study was conducted to evaluate the clustering capabilities of K-Means and HDBSCAN. The overall metrics results show that K-Means provides the best results for all metrics used. Despite the lightweight of our model (3,6 GFLOPs), it achieved a performance comparable to other state-of-the-art approach.

Keywords: Embedding learning, Unsupervised Clustering, Breast Cancer images, ArcFace loss, lightweight clustering method

INTRODUCTION

Several types of cancer can affect humans, among which breast cancer is one of the most prevalent, primarily affecting women. Currently, this disease represents a major public health challenge worldwide, particularly in Africa, where the number of new cases and deaths continues to rise. According to the GLOBOCAN 2020 program, breast cancer is the most common cancer among women in Africa, with approximately 186,598 new cases and 85,787 deaths recorded in 2020 [16], [4]. This high mortality rate is generally attributed to late diagnosis, limited access to screening and treatment facilities, and significant inequalities within healthcare systems [5], [18].

In Madagascar, data on breast cancer remain very limited, which constitutes a significant constraint for monitoring the progression of the disease. However, according to estimates from the World Health Organization, breast cancer was responsible for approximately 629 deaths in 2020, representing 0.38% of all deaths in the country, with an age-standardized mortality rate of 8.32 per 100,000 inhabitants [20], [4]. Survival rates remain

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

low, particularly for patients diagnosed at advanced stages. Therefore, early detection is crucial to reducing mortality associated with this disease.

In recent years, artificial intelligence has provided innovative solutions across various domains, including healthcare, particularly for the early detection of diseases. For instance, in [1], a comparative study on breast cancer classification was conducted using transfer learning, meta-learning, and ensemble learning techniques. The Breast Ultrasound Imaging (BUSI) dataset was used to evaluate the performance of the proposed methods, and the results demonstrated that meta-learning achieved superior performance. In [9], a method based on wavelet and curvelet coefficients was proposed to obtain multi-resolution representations of images; features extracted at each resolution level were then used as input to a binary tree classifier.

In [14], a clustering ensemble approach was introduced, combining multiple algorithms to improve performance. Similarly, the work presented in [7] proposed a hybrid clustering framework that integrates clustering with other machine learning techniques such as classification and regression. An example of such an approach is the Deep Clustering Ensemble (DCE) [11], which combines deep learning with ensemble strategies. In [19], Graph Convolutional Clustering (GCC) was proposed, integrating graph convolutional networks with clustering to enhance the efficiency and accuracy of clustering for graph-structured data. For large-scale data, the Distributed Deep Clustering (DDC) framework [21] leverages distributed computing platforms such as Apache Spark to enable efficient training of deep clustering models across multiple machines.

Although these approaches achieve high performance, they are often complex and computationally intensive. In contrast, this work proposes a lightweight clustering method based on the combination of MobileNetV3Small, ArcFace, and clustering algorithms such as K-Means and HDBSCAN.

The remainder of this paper is organized as follows: Section II reviews related work on breast cancer classification, Section III describes the proposed approach, Section IV presents the results and discussion, and Section V concludes the paper.

RELATED WORK

The present work focuses on the breast cancer images clustering using images dataset. Several works have already been carried out in this direction. Various approaches were existed to cluster these ones. The authors proposed in [15] a comparative study of three unsupervised algorithms for the task of breast Magnetic Resonance (MRI) lesion segmentation, namely, Gaussian Mixture Model clustering, K-Means clustering and a marker-controlled Water shed transformation-based method. They used a dataset MR image acquired from 1.5 T scanners Magnetom Avanto. and 3.0 T Magnetom Verio, Siemens Healthineers, Erlangen, Germany, with dedicated breast array coils and the patient in a prone position. they applied to the all methods on breast MRI slices following selection of regions of interest (ROIs) by an expert radiologist. They used the Dice similarity coefficient (DSC), Jaccard index (JI), Hausdorff distance and precision recall metrics to evaluated the segmentation accuracy. The result shows that the marker-controlled Watershed transformation achieved higher segmentation accuracy compared to the other algorithm.

In [2], the authors proposed an automatic approach to segment masses in mammograms. Their method used hierarchical clustering to isolate the salient area, and then features are extracted to reject false detection. They used two datasets (mini-MIAS and DDSM) to evaluate their approach. Support Vector Machine (SVM) is used as classifier. The total accuracy of the system is 83,43%. The result shows that their method is efficient compared with other techniques.

The Work presented in [13] made comparative studies of two clustering methods: K-means and K-medoids clustering or Partitioning Around Medoids (PAM). In their approach, no pretrained model and label were need. The dataset consisted of 458 benign which is 65.5% of the dataset and 241 malignant which is also 34.5% of the dataset. They used Weka software to pre-process the data. Hopkins Statistic is used to evaluate the clustering

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

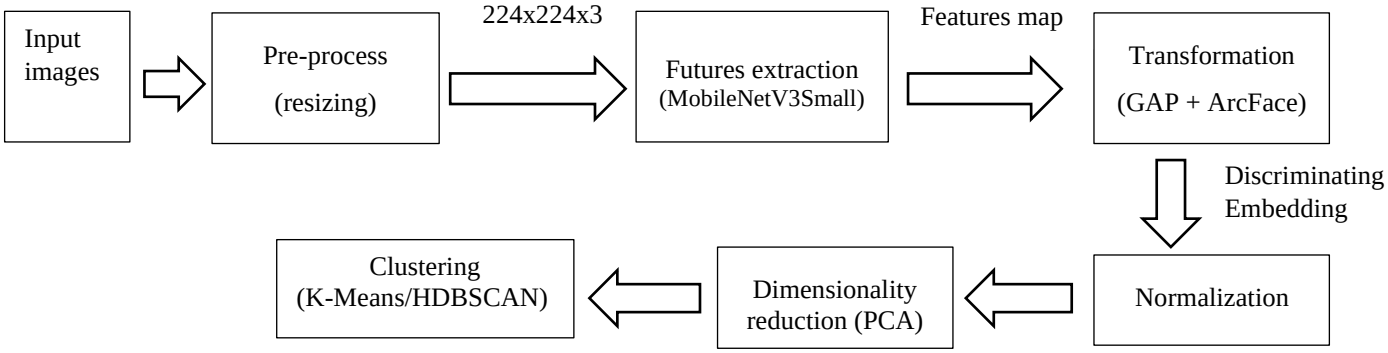
quality of the dataset and silhouette width is used to measure the performance of the clustering methods. The experiment results show that k-Means achieved slightly higher silhouette width (0.58) compared with PAM (0.57).

In [10], authors conducted a comprehensive comparison of three clustering methods: K-Means, Agglomerative, and Gaussian Mixture Models applied to breast cancer dataset downloaded from Kaggle. They evaluate the performance of these clustering techniques using the Silhouette Score, Calinski-Harabasz Score, and Davies Bouldin Score. The results show: Silhouette Score k-Means 0.4711, Agglomerative 0.4631 and GMM 0.4703; Calinski-Harabasz Score: k-Means 523.4070, Agglomerative 494.1320 and GMM 429.3527; Davies-Bouldin Score kmeans 0.9263, Agglomerative 0.9750 and GMM 1.0203. After analysis, K-Means performs the best in terms of creating distinct, well-separated, and compact clusters for the breast dataset.

MATERIALS AND METHODS

Our Approach

In this section, we describe our approach to cluster breast cancer ultrasound images. Dataset publicly available in Kaggle web site [22] was used to evaluate the model. The data is injected to the backbone for extracting the features map. This one is fed to several pool layer to product features vector. This vector is then projected into a low-dimensional latent space to form an embedding. To improve data separability in this space, an ArcFace-type loss function [6] is applied, imposing an angular constraint between classes and promoting the formation of discriminating embeddings. Finally, these embedding are exploited by unsupervised clustering algorithms, after having applicated normalization and reduction of dimensionality, such as K-Means or HDBSCAN, in order to automatically clustering images according to their morphological similarities, in particular for the distinction between benign and malignant lesions. Figure 1 illustrate our approach.



GAP: Global Average Pooling

Figure 1: Our approach

Dataset Description

For this study, we used a publicly available dataset downloaded from the Kaggle website. The dataset comprises pre-labelled ultrasound images classified into two categories: benign and malignant breast cancer cases. The original images had been previously augmented through rotation and sharpening transformations, resulting in a total of 9,016 images. These images were already partitioned into training and validation sets.

The training set includes 4,074 benign and 4,042 malignant images, while the validation set consists of 500 benign and 400 malignant images. It should be noted that the image resolutions vary across the dataset.

For the purposes of this study, all images were consolidated into a single directory, thereby removing their original label-based folder structure. Subsequently, a down sampling technique was employed to ensure class balance. Following this process, each class contained 3,000 images.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

To ensure compatibility with the feature extractor, all images were resized to 224 × 224 pixels prior to being input into the model for clustering into benign or malignant categories. No additional pre-processing steps were required. Figure 2 presents representative samples from the dataset.

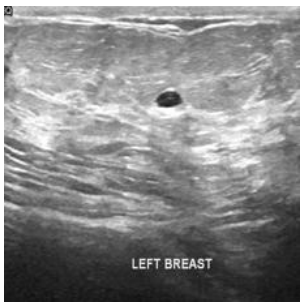


Figure 2.a: benign case

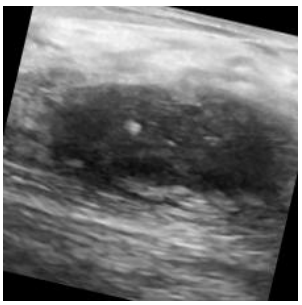


Figure 2.b: malignant case

Proposed Method

The proposed model employs a hybrid approach combining supervised deep learning and metric learning to produce discriminating vector representations suitable for classification and clustering tasks. The architecture relies on the MobileNetV3Small convolutional network as a feature extractor. This model, known for its computational efficiency and lightweight nature, is used without its final classification layer to retain only its ability to generate relevant visual descriptors. The input images are resized and then fed into the network. The pre-trained weights are initially frozen, thus leveraging generic representations learned from large datasets.

At the output of the convolutional backbone, a projection head is added to transform the feature maps into a compact embedding vector. This model head consists of a global pooling layer followed by several fully connected layers, incorporating nonlinear activation functions and Batch Normalization. The goal of this structure is to project the data into a low-dimensional (64-dimensional) latent space, where the discriminating properties between classes are strengthened. Furthermore, data augmentation techniques, including random geometric transformations such as horizontal flips, rotations, and zooms, are applied during training to improve the model's generalization ability to different data.

The originality of this approach lies in the integration of the ArcFace layer, which does not merely replace a conventional loss function, but redefines the geometry of the learning space. Unlike a standard softmax loss, ArcFace imposes an explicit angular constraint by normalizing both the embeddings and the classification weights, and then introducing an additive angular margin for the target class. This formulation transforms the classification problem into learning on a hypersphere, where only angular distances are relevant.

Within this model, ArcFace assumes two fundamental and complementary roles. On the one hand, it acts as a highly discriminative supervised loss function by reducing intra-class variance (compactness) and increasing inter-class distance (separability). On the other hand, it explicitly structures the embedding space so that it becomes intrinsically compatible with unsupervised clustering methods. Indeed, by constraining the data to organize into angularly separated clusters, ArcFace produces a latent space where class boundaries naturally emerge, even in the absence of labels during the clustering stage. In other words, although the training is supervised, the geometry induced by ArcFace promotes direct usability in an unsupervised context.

The training strategy proceeds in two complementary phases. In the first phase, the convolutional backbone is frozen, and only the projection layers as well as the ArcFace layer are trained. This stabilizes the learning of embeddings without disrupting the general representations already learned. In the second phase, partial fine-tuning is performed by unfreezing the last layers of the backbone, with a reduced learning rate, in order to more precisely adapt the features to the specifics of the dataset. This progressive approach balances stability and specialization.

Once the model is trained, only the embedding component is retained to project images into the learned latent space. These vectors are then normalized (consistent with ArcFace's angular constraint), and subsequently

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

subjected to dimensionality reduction using the Principal Component Analysis (PCA) algorithm. Beyond simple compression, the use of PCA offers several advantages in this context. First, it removes noise and residual correlations between dimensions, improving the overall structure of the space. Second, by concentrating discriminative information in the principal components, it enhances cluster separability for algorithms sensitive to Euclidean distance such as K-Means. Third, it reduces the effective dimensionality, mitigating the effects of the "curse of dimensionality" and stabilizing similarity metrics.

The t-SNE map was computed using PCA components. It shows the natural clustering of the data. The resulting embeddings are then used in unsupervised clustering methods, notably K-Means and HDBSCAN. The optimal number of clusters for K-Means is automatically determined from the silhouette coefficient (see figure 3), while HDBSCAN detects structures of varying density while identifying points considered noise.

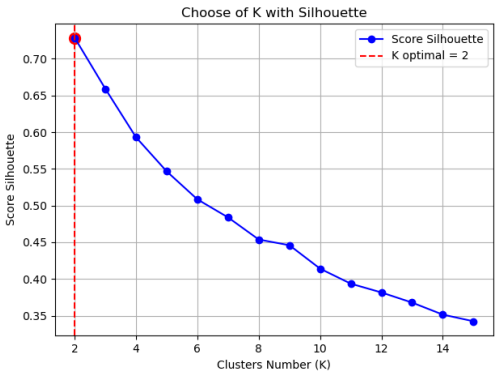


Figure 3: search k values using Score silhouette

t-SNE maps of the embedded representation is shown in Figure 4.

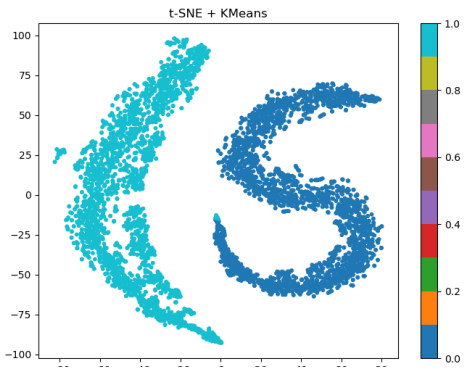


Figure 4.a: K-means t-SNE map for PCA components

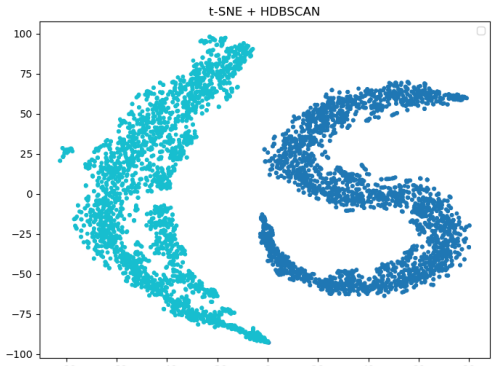


Figure 4.b: HDBSCAN t-SNE map for PCA components

RESULTS AND DISCUSSION

The proposed model was evaluated on a dataset of histopathological images of breast cancer, composed of two balanced classes: benign and malignant cases. The images, initially stored without class organization.

Performance evaluation is carried out using standard classification and clustering metrics. Accuracy measures the proportion of correctly predicted labels among all samples. The F1-score [3], defined as the harmonic mean of precision and recall, provides a balanced assessment of classification performance, particularly in the presence of class imbalance.

However, in the context of clustering, both accuracy and F1-score require an explicit mapping between cluster assignments and ground truth labels. This mapping introduces a dependency on label alignment and may bias the evaluation, as clustering algorithms are inherently permutation-invariant. Consequently, these metrics should be interpreted with caution, as they do not purely reflect the intrinsic quality of the clustering structure but rather the quality of the post-hoc label matching.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

To address this limitation, clustering performance is further evaluated using permutation-invariant metrics such as the Adjusted Rand Index (ARI) [8] and Normalized Mutual Information (NMI) [17]. The ARI quantifies the similarity between two partitions by correcting the Rand Index for chance, yielding values in the range [-1, 1], where 1 indicates perfect agreement Comparing Partitions. The NMI is an information-theoretic measure based on mutual information, normalized to ensure values between 0 and 1, where higher values indicate better agreement between clustering assignments and ground truth labels Elements of Information Theory.

To ensure robustness and stability of the results, K-Fold cross-validation [12] (k=10), as formalized by Ron Kohavi, is employed. In this procedure, the dataset is partitioned into ten folds, where each fold is used once as a validation set while the remaining folds are used for training. This approach enables reliable performance estimation by reducing variance associated with a single train-test split, especially when evaluating learned embeddings for unsupervised clustering tasks.

The results obtained with the K-Means algorithm show high overall performance, with a mean accuracy of 0.945 and a similar mean F1 score. Table I.a and Table I.b show the predictive results of K-Means and HDBSCAN. The low variability observed between the different folds demonstrates the robustness of the model and its ability to produce stable representations. In addition, the clustering metrics indicate good agreement between the predicted clusters and the actual classes, with a mean Adjusted Rand Index of 0.792 (± 0.035) and a Normalized Mutual Information of 0.695 (± 0.040). These results suggest that the learned embedding space exhibits a coherent structure, enabling efficient separation of the two classes.

Evaluation using the HDBSCAN algorithm reveals slightly lower performance, with an average accuracy of 0.9255 and an F1 score of 0.9258. The ARI (0.702 ± 0.032) and NMI (0.618 ± 0.028) values confirm this trend. This difference can be explained by the very nature of HDBSCAN, which relies on a density-based approach and is particularly sensitive to the distribution of data in the latent space. In this case, the embeddings appear to form relatively compact and well-separated groups, a configuration more favourable to partitioning methods like K-Means than to density-based clustering approaches.

These results highlight the effectiveness of the adopted strategy, particularly the use of ArcFace loss to structure the embedding space. By imposing an angular constraint between classes, this approach promotes strong intra-class compactness and marked inter-class separation, which directly translates into better clustering performance. The relatively small difference between accuracy and F1 score also indicates a good balance in the classification of the two classes, which is consistent with the balanced distribution of the dataset.

Finally, it is worth noting that the proposed approach transforms an initially supervised problem into an unsupervised representation space structuring task. The results demonstrate that the learned embeddings are sufficiently discriminating to allow for efficient class separation without direct recourse to a supervised classifier during the inference phase. This property opens up interesting possibilities, particularly for applications where annotations are limited or partially available, as is often the case in medical imaging.

Table I.a: K-Means predictive results

Validation Accuracy	F1 Score
0.9350	0.9350
0.9450	0.9450
0.9567	0.9567
0.9467	0.9467
0.9517	0.9517
0.9350	0.9350
0.9617	0.9617
0.9500	0.9500
0.9283	0.9283
0.9400	0.9400

Table I.b: HDBSCAN predictive results

Validation Accuracy	F1 Score
0.9217	0.9220
0.9317	0.9320
0.9250	0.9253
0.9250	0.9254
0.9317	0.9320
0.9217	0.9219
0.9267	0.9268
0.9383	0.9384
0.9233	0.9234
0.9100	0.9108

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

In summary, the performance obtained confirms the relevance of the proposed architecture and the adopted learning strategy, while highlighting the interest of metric-based learning approaches for classification and clustering tasks in a biomedical context. Furthermore, compared with previous state-of-the-art works, our approach achieves better performance than [2] and [3] in terms of both accuracy and silhouette score.

CONCLUSION AND PERSPECTIVES

The present work was focused on breast cancer images clustering using hybrid approach combining transfer learning, ArcFace and unsupervised clustering. Our approach was tested with ultrasound breast cancer images. Despite the encouraging performance achieved, several limitations must be highlighted to accurately assess the scope of the results and identify areas for improvement. First, although the dataset is balanced between benign and malignant classes, it remains limited to a relatively simple binary configuration. This constraint reduces the inherent complexity of the problem and prevents a full evaluation of the model's ability to generalize to more realistic scenarios, particularly those involving a greater number of histopathological subtypes or unbalanced class distributions, which are common in clinical practice.

The evaluation based primarily on clustering methods applied to embeddings. While metrics such as the Adjusted Rand Index and Normalized Mutual Information provide a relevant indication of the quality of latent space structuring, they do not replace a comprehensive evaluation within a strictly supervised framework. In particular, the lack of direct comparison with standard classifiers (e.g., a classic softmax layer) makes it difficult to estimate the actual gain from using ArcFace loss in this specific context.

Furthermore, the choice of hyperparameters, especially those related to the ArcFace layer (angular margin and scaling factor), has not been systematically studied. However, these parameters play a crucial role in structuring the embedding space and can significantly influence performance.

In terms of future directions, several avenues for improvement can be explored. First, extending the study to more diverse datasets, including multiple classes and from different sources, would allow for a better assessment of the model's robustness and generalizability. Integrating domain adaptation techniques or pre-training specific to the medical domain could also improve the quality of the learned representations.

Furthermore, a more in-depth exploration of metric learning approaches is a promising avenue. Variants of ArcFace loss or other distance-based loss functions (such as Triplet Loss or Contrastive Loss) could be investigated to compare their impact on the structuring of the latent space. Similarly, the introduction of attention mechanisms or more recent models could enable the capture of finer and more discriminating features.

Finally, a particularly interesting prospect lies in integrating the model into a semi-supervised or weakly supervised framework, where only a portion of the data is annotated. In this context, the model's ability to produce discriminating embeddings could be leveraged to improve overall performance while reducing the annotation cost, a major challenge in medical imaging. The combined use of clustering and active learning techniques could also be a promising way to progressively refine annotations and improve system performance.

In conclusion, although the results obtained are encouraging, they constitute an intermediate step towards the development of more robust, generalizable systems adapted to the real constraints of biomedical applications.

REFERENCES

1. Ali et al. (2023). Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks. Diagnostics 2023, volume 13. 19 pages, 2023. https://doi.org/10.3390/diagnostics13132242
2. Bilal Ahmed Lodhi (2021). Unsupervised Method to Localize Masses in Mammograms. arXiv: 1904.06044v1[cs.CV] 12 Apr 2019.IEEE Access, vol. 9, pp. 99327-99338, 2021.
3. C. D. Manning, P. Raghavan, and H. Schütze (2008). Introduction to Information Retrieval. Cambridge University Press, 2008.
4. Ferlay, J., Ervik, M., Lam, F., et al. (2021). Global Cancer Observatory: Cancer Today. International Agency for Research on Cancer (IARC), 2021.

5. Jedy-Agba, E., McCormack, V., Adebamowo, C., & dos-Santos-Silva, I. (2016). Stage at diagnosis of breast cancer in sub-Saharan Africa: A systematic review. The Lancet Global Health, 4(12), e923–e935, 2016.
6. Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou (2015). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. journal of latex class files, vol. 14, no. 8, august 2015. arXiv:1801.07698v4 [cs.CV] 4 Sep 2022
7. Kashef, R., & Kamel, M. S. (2009). Cooperative clustering. Pattern Recognition, 42(10), pp. 2324-2349, 2009.
8. L. Hubert and P. Arabie (1995). Comparing Partitions, Journal of Classification, vol. 2, no. 1, pp. 193–218, 1985. https://doi.org/10.1007/BF01908075
9. M. M. Eltoukhy and I. Faye (2014). An optimized feature selection method for breast cancer diagnosis in digital mammogram using multiresolution representation. Appl. Math, 8(6), pp. 2921–2928, 2014.
10. Nikhil Sanjay Suryawanshi (2023). Enhancing Breast Cancer Diagnosis Through Clustering: A Study of KMeans, Agglomerative, and Gaussian Mixture Models. International Journal of Innovative Science and Research Technology, Volume 8, Issue 7, pp 3497-3504, July 2023.
11. Peng, X., et al. (2017). Deep clustering via integrating sparse subspace clustering analysis and deep representation. Pattern Recognition Letters, 98, pp. 74-83, 2017.
12. Ron Kohavi (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995.
13. Somenath Chakraborty (2021). Beddhu Murali. Investigate the Correlation of Breast Cancer Dataset using Different Clustering Technique. arXiv:2109.01538v1[cs.CV], 2021. https://doi.org/10.48550/arXiv.2109.01538
14. Strehl, A., & Ghosh, J. (2002). Cluster ensembles a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research,3(Dec), 583-617, 2002.
15. Sulaiman Vesal, Nishant Ravikumar, Stephan Ellman, Andreas Maier (2018). Comparative Analysis of Unsupervised Algorithms for Breast MRI Lesion Segmentation. arXiv: 1802.08655v1[cs.CV] 23 feb 2018. 6 pages, 2018. https://doi.org/10.48550/arXiv.1802.08655.
16. Sung, H., Ferlay, J., Siegel, R. L., et al. (2021). Global cancer statistics 2020. CA: A Cancer Journal for Clinicians, 71(3), pp. 209–249, 2021.
17. T. M. Cover and J. A. Thomas (2006). Elements of Information Theory, 2nd ed., Wiley-Interscience, 2006.
18. Vanderpuye, V., Grover, S., Hammad, N., et al. (2017). An update on the management of breast cancer in Africa. Infectious Agents and Cancer, 12(13), 2017.
19. Wang, Y., et al. (2023). Graph Convolutional Clustering: A Deep Learning Approach to Graph Clustering. In Proceedings of the 13th ACM International Conference on Web Search and Data Mining, pp. 861-869, 2023.
20. World Health Organization (WHO). (2020). Global Health Observatory data repository. 2020.
21. Zaheer, M., Reddi, S., Sachan, D., Kale, S., & Kumar, S. (2019). Distributed Deep Clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9489-9498, 2019.
22. https://www.kaggle.com/datasets/vuppalaadithyasairam/ultrasound-breast-images-for-breast-cancer