Applied Deep Learning Architectures for Breast Cancer Screening Classification

Asma Zizaan[1], Ali Idri[1,2] and Hasnae Zerouaoui[1]
[1]*Modelling, Simulation and Data Analysis, Mohammed VI Polytechnic University, Benguerir, Morocco*
[2]*Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat, Morocco*

Keywords: Computer-Aided Screening, Deep Convolutional Neural Networks, Transfer Learning, Breast Cancer Screening, Mammogram Classification, Image Processing.

Abstract: Breast cancer (BC) became the most diagnosed cancer, making it one of the deadliest diseases. Mammography is a modality used for early detection of breast cancer. The objective of the present paper is to evaluate and compare deep learning techniques applied to mammogram images. The paper conducts an experimental evaluation of eight deep Convolutional Neural Network (CNN) architectures for a binary classification of breast screening mammograms, namely VGG16, VGG19, DenseNet201, Inception ResNet V2, Inception V3, ResNet 50, MobileNet V2 and Xception. This evaluation was based on four performance metrics (accuracy, precision, recall and f1-score), as well as Scott Knott statistical test and Borda count voting system. The data was extracted from the CBIS-DDSM dataset with 4000 images. And results have shown that DenseNet201 was the most efficient model for the binary classification with an accuracy of 84.27%.

1 INTRODUCTION

One of the most diagnosed types of cancer among women is breast cancer, with statistics showing that one out of eight females will be diagnosed with breast cancer in their lifetime (The American Cancer Society medical and editorial content team, 2022). In 2020, breast cancer was reportedly diagnosed in 2.3 million women and has caused 685 000 deaths globally (Breast Cancer, n.d.). Survival rates started to appear promising in countries where early detection programs were combined with multiple treatment options to eradicate this invasive illness (Coleman, 2017). Breast cancer can be detected early through screening mammography, which is one of the most common screening modalities of our time.

Medical image processing is defined as the use and the investigation of image files of the human body, typically collected from a Computed Tomography (CT), Magnetic Resonance Imaging (MRI) scanner, or another type of X-ray system using computerized quantification and visualization tools. The purposes of this analysis are to help search for or diagnose disorders and guide medical procedures such as surgery planning and treatments (McAuliffe et al., 2001). Machine learning, specifically deep learning, has sparked major interest in its application to medical image processing due to its rapid progress. Various previously published works applied deep learning models to multiple medical fields such as breast cancer and diabetic retinopathy (Lahmar & Idri, 2022; Zerouaoui & Idri, 2022). For example, (Zizaan & Idri, 2022): Shen et al. has constructed a deep learning algorithm that may accurately identify breast cancer instances on routine screening mammograms, using an "end-to-end" training strategy. This study showed promising results and is trained to reach a high accuracy when applied to similar mammogram datasets (Shen et al., 2019). In the same manner, Agarwal et al. proposed a CNN framework for automated mass detection in full-field digital mammograms (FFDM) using VGG16, Resnet50, and InceptionV3 (Agarwal et al., 2019).

The common downfall to the mentioned articles is: (1) the lack of variety of the used DL architectures for better comparison, and (2) the choice of evaluation methods that is often limited to the accuracy and area under curve (AUC).

Thus, the aim of this paper is to develop as well as evaluate various DL techniques applied to the binary classification of the CBIS-DDSM dataset of BCS mammograms, namely VGG16, VGG19, DenseNet201, Inception ResNet V2, Inception V3,

Zizaan, A., Idri, A. and Zerouaoui, H.
Applied Deep Learning Architectures for Breast Cancer Screening Classification.
DOI: 10.5220/0011723700003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 617-624
ISBN: 978-989-758-623-1; ISSN: 2184-433X

ResNet 50, MobileNet V2 and Xception architectures. The empirical evaluation is set over two steps, four performance measures in the first place, namely: accuracy, precision, recall, f1-score, and secondly, statistical testing using the Scott-Knott test and Borda count system in order to select the best performing model out of the eight. Such evaluation methods are used to compare, cluster ad rank DL models (Idri et al., 2016; Ottoni et al., 2019; Zerouaoui et al., 2021; Zerouaoui & Idri, 2022). The results of this study are discussed over two research questions (RQs):

(RQ1): What is the overall performance of DL techniques in BCS binary classification?

(RQ2): Are there any DL techniques that noticeably outperform the others?

Accordingly, the main contributions of this article are: (1) the development of eight end-to-end CNN architectures, (2) preparing the data by using the Contrast Limited Adaptive histogram equalization (CLAHE) (Pizer et al., 1987) technique and intensity normalization, and finally (3) comparing the results of each model and selecting the best performing among them.

The following is a breakdown of the paper's structure: Section 2 presents the related works. In Section 3, lays out the background of the study. As for section 4, it explains the data preparation process as well as the configuration of the eight DL techniques. Section 5 presents and discusses the empirical results. And finally, section 6 outlines the conclusions and future works.

2 RELATED WORKS

This section summarizes the findings of previous studies investigating deep learning techniques for BCS classification. Zizaan et al. published a systematic mapping study (SMS) of the use of machine learning in BCS (Zizaan & Idri, 2022). A total of 66 papers published between 2011 and 2021 were selected for analysis, the main findings of the SMS were:

(1) The most common publication venue is journals (58.88%), followed by conferences (6.9%), and books (2.3%). These publications showed a peak in 2019.

(2) The most frequent paper type is evaluation research (92.4%) including 31.8% of solution proposal papers. And the least frequent articles are reviews (7.5%).

(3) The most used BCS modality is mammography (61%), followed digital breast tomosynthesis, MRI, and ABUS imaging (10%, 7%, and 7% respectively).

Out of the sixty-six selected papers, nine articles used the DDSM dataset. Table 1. presents the five most recent studies and their findings. The most used DL techniques were VGG16 and Resnet50, while the

Table 1: Overview of five related studies.

Authors	DL techniques	Performance metrics	Results
Shen et al.	VGG16, Resnet50, Hybrid networks	Accuracy, AUC	On several mammography platforms, automatic deep learning methods can be easily trained to achieve excellent accuracy.(Shen et al., 2019).
Aboutalib et al.	AlexNet	AUC	The DDSM dataset had the overall best performance. This could be as a result of the higher dataset size or the features of the dataset itself. (Aboutalib et al., 2018).
Chougrad et al.	ImageNet, VGG16, RESNET50, and inceptionv3	AUC, Accuracy	Achieve 97.35% accuracy as well as 0.98 AUC on the DDSM dataset (Chougrad et al., 2018)
Saranyaraj et al.	Deep Convolutional neural network (DCNN), ResNet, GoogleNet and VGG	Accuracy, specificity, AUC, recall, precision, and F1-score	Increase the mean classification accuracy to 97.46% and the overall classification accuracy to 96.23% (Saranyaraj et al., 2020)
Agarwal et al.	VGG16, ResNet50, and InceptionV3.	Accuracy	Develop an automated framework that has obtained the best results based on TPR and FPI (Agarwal et al., 2019)

most common performance metrics are accuracy and area under curve (AUC). Some of the noticeable limitations were the low number of applied DL techniques, averaging at three architectures, as well as the performance metrics and the lack of statistical tests.

3 EXPERIMENT CONFIGURATION

This section presents the data pre-processing tasks, the parameter tuning of the DL models, as well as the empirical design.

3.1 Data Preparation

This dataset consists of images from CBIS-DDSM dataset (Eric A. Scuccimarra, 2018). Digital Imaging and Communications in Medicine format (DICOM) pictures totaling 10,239 were collected from 1,566 patients across 6,775 trials to build this dataset (Lee et al., 2017). Some Pre-Processing tasks have already been done such as extracting the region of interest (ROIs) and resizing to 299x299. The primary dataset contains 55890 training mammograms, with 7825 (14%) positive cases and 48064 (86%) negative cases.

Considering that more than half of the images in the CBIS-DDSM dataset were labelled negative, which is considered as a limitation, resampling the images from the DDSM dataset was made by using data augmentation. It is worth mentioning that the CBIS-DDSM dataset is a subset of the DDSM dataset, meaning that the images that were sampled from DDSM are of the same nature as the images existing in the CBIS-DDSM dataset.

Further image processing was made using the Contrast Limited Adaptive histogram equalization (CLAHE) technique to enhance the images' contrast as well as intensity normalization. In Figure 1, the data preparation process is detailed: First, acquiring the data from the DDSM and CBIS-DDSM datasets with the same number of samples in each class. Then, enhancing the contrast of the images using the CLAHE technique. Lastly, applying the min-max normalization as shown in Equation 1 to the input photos will normalize them to the conventional distribution. Results of the image enhancing technique are shown in Figure 2.

$$. x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

3.2 Model Configurations

As the present study is a binary classification of BCS mammogram data extracted from the publicly available dataset CBIS-DDSM, eight different end-to-end CNN architectures were implemented using several parameter tuning experiments. To train the models after the data preparation phase was finished, the transfer learning technique served as a practical tool to import models pre-trained in the ImageNet dataset (Fei-Fei et al., 2010).

To summarize the parameter tuning, the batch size was set to 32 and the number of epochs to 200. As for the optimization, the output layer optimization function was SoftMax, and the optimizer Adam (adaptive moment estimation) was chosen (Kingma & Ba, 2015) with an initial learning set to 0. 00001. Moreover, the loss function was set as the cross entropy, and all the layers were frozen for the transfer learning.

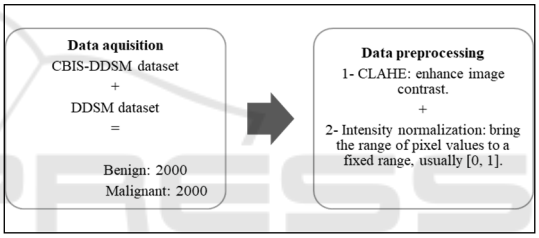

Figure 1: Data preparation process.

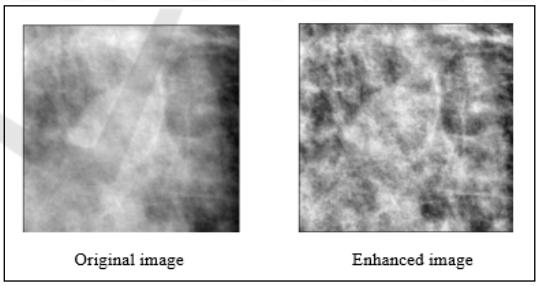

Figure 2: Example of CLAHE image transformation.

3.3 Data Splitting and Performance Metrics

In this study, we used K-fold cross validation with k equal to 5 to apply and evaluate the DL models. We also provided the performance metrics' average for each DL method. Four metrics—accuracy, precision, recall, and f1-score—were used to evaluate the performance of the eight DL classifiers. These metrics are provided by Equations 2 through 5.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-score = 2 \times \frac{Recall \times Precision}{Recall+Precision} \quad (5)$$

where:

- TP: true positives.
- FP: false positives.
- TN: true negatives.
- FN: false negatives.

In addition to the four performance metrics, two statistical tests were used:

Scott Knott test is a type of exploratory clustering techniques that is commonly employed in the context of analysis of variance (ANOVA). Scott and Knott proposed it in 1974 as a way to distinguish overlapping groups using numerous comparisons of treatment means (Jelihovschi et al., 2014).

Borda count is a voting method for single winner election methods. Candidates are given points based on their ranking in this method: 1 point for last choice, 2 points for second-to-last choice, and so on. The entire point totals for all ballots are added together, and the candidate with the highest point total wins (García-Lapresta & Martínez-Panero, 2002).

3.4 Experimental Design

The methodology used to conduct the empirical evaluations consists of three steps as shown in Fig. 3.

(1) Assess four performance metrics of each variant of the deep learning architectures (VGG16, VGG19, DenseNet201, Inception ResNet V2, Inception V3, ResNet 50, MobileNet V2 and Xception) as well as a CNN baseline.

(2) Select the DL architectures that outperformed the CNN baseline by comparing the accuracy results.

(3) Use the Skott Knott test to build clusters of the selected DL models and select the best SK cluster.

(4) To choose the ideal DL architecture, rank the top SK cluster using the Borda count voting technique based on the four performance criteria (accuracy, precision, recall, and F1-score).

It is worth noting that comparable approaches were utilized in previous works. (Azzeh et al., 2017 ; Idri et al., 2018 ; Idri & Abnane, 2017 ; Worsley, 1977 ; Zerouaoui et al., 2021)

4 RESULTS AND DISCUSSION

This section presents and discusses the evaluations' findings for the eight DL approaches used on the CBIS-DDSM dataset. The performances of the DL techniques were evaluated using four numerical performance metrics: accuracy, recall, precision, and F1-score, as well as two statistical tests: Borda count, and Scott-Knott. First, the performance of each DL approach is examined, and the ones that outperform the CNN baseline in accuracy are chosen. (RQ1).

Then, we utilize the Borda count voting mechanism to rank the DL approaches belonging to the best SK cluster after using the SK statistical test based on accuracy to cluster the chosen DL techniques (i.e., accuracy greater than the CNN baseline model) (RQ2).

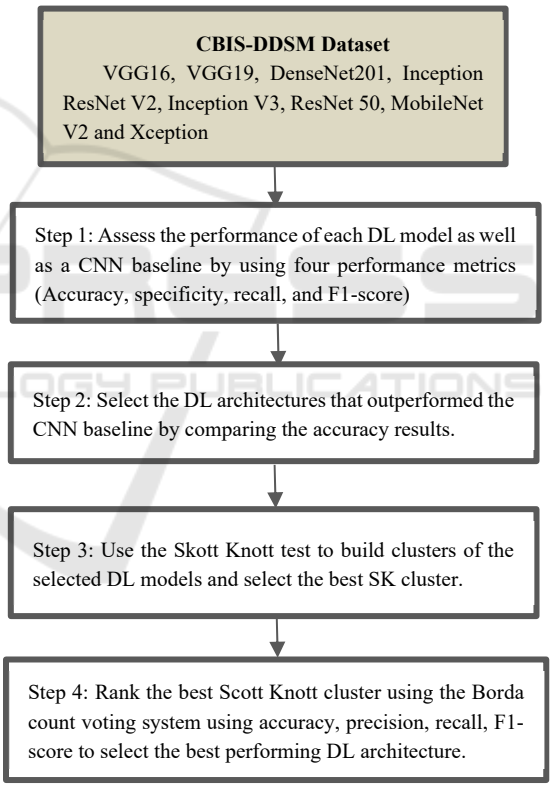

Figure 3: Experimental design.

4.1 Assessment of the DL Models' Performance

Eight different CNN architectures which were previously pre-trained on ImageNet (VGG19, VGG16, ResNet50, Inception V3, DenseNet201, MobileNet V2, Xception, and Inception ResNet V2) are applied over the DDSM dataset in Python using

Tensorflow (Abadi et al., 2016), Keras, SciKit-Learn (Pedregosa et al., 2011), Pandas(Reback et al., 2022), Matplotlib (Hunter, 2007), NumPy, and Seaborn (Waskom, 2021) frameworks, on a tensor processing unit (TPU) provided by Google in collab notebook.

Figure 4 and Table 2 show the accuracy values of the DL models as well as the CNN baseline. These results are summarized in Table 2 and reveal that:

a) DenseNet201 unlocks better performance than the other CNN architectures with 84.27% accuracy and 84.27% F1 score.
b) ResNet50 was the weakest model showing a noticeably lower accuracy (76.35%) and F1-score (76.46%).
c) MobileNetV2 outperforms Inception V3 by a 2.67% higher accuracy.

Table 2: Performance metrics.

CNN Architecture	Accuracy	Precision	Recall	F1 Score
CNN baseline	82.07%	80.98%	84.25%	82.50%
VGG19	81.00%	80.76%	81.85%	81.14%
VGG16	82.82%	81.48%	85.19%	83.26%
InceptionV3	80.75%	80.77%	81.10%	80.87%
DenseNet201	**84.27%**	**84.55%**	**84.0%**	**84.27%**
MobileNetV2	83.42%	83.47%	83.55%	83.48%
Xception	82.57%	82.64%	82.60%	82.61%
InceptionResnetV2	83.12%	83.34%	82.65%	83.14%
ResNet50	76.35%	76.27%	76.92%	76.46%

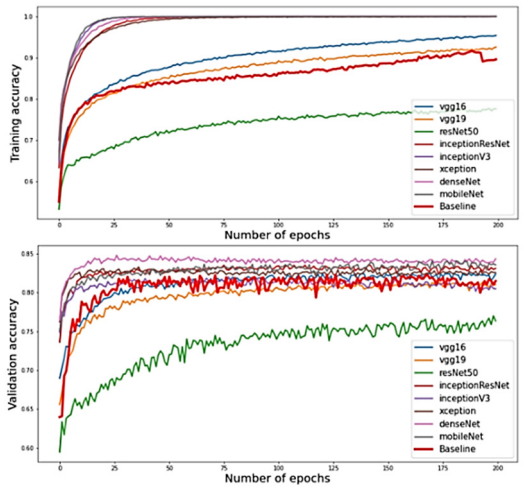

Figure 4: Validation and training accuracy values of each DL model.

Hereafter, the models to be discussed are the ones that have scored a better performance than the CNN baseline, namely: VGG16, DenseNet201, MobileNetV2, Xception, and InceptionResNetV2.

4.2 Selection of the Best Performing DL Model

Based on the results obtained, and to guarantee a more precise selection of the best group of architectures, further testing was required. In particular, the hierarchical clustering of Scott Knott was applied to the DL architectures which scored a higher accuracy than the CNN baseline.

The outcome of the SK test shows that the selected DL models all belong to a single cluster (figure 5). As a result, these four DL approaches have similar prediction skills in terms of accuracy values. This cluster contains DenseNet201, MobileNet V2, InceptionResNet V2, Xception, and VGG16.

Finally, selecting the best model from the Scott Knott (table 3) cluster has been made by the Borda count voting system taking into consideration the four metrics (accuracy, recall, precision, and f1-score).

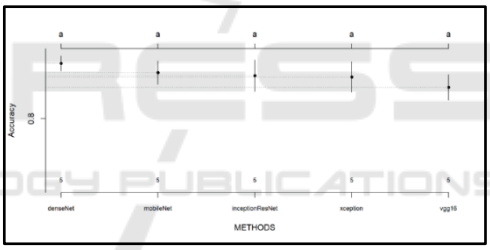

Figure 5: Results of Scott Knott test of the DL techniques.

Table 3: Borda count rankings.

DL Model	Rank
DenseNet201	1
MobileNet V2	2
VGG16	3
Inception ResNet V2	4
Xception	5

5 THREATS TO VALIDITY

This section discusses the threats to the validity of this paper from both external and internal perspectives.

5.1 Internal Validity

In order to strengthen the robustness of the mean accuracy of the eight DL designs, this work applied

the K-fold cross validation approach. When it comes to optimization, the Adam (adaptive moment estimation) optimizer's superior performance and quick learning rate convergence make it preferable to the traditional stochastic gradient descent method (Zhang, 2019).

5.2 External Validity

The CBIS-DDSM dataset, which provides screening mammography images, was utilized in this investigation; however, we are unable to extend the findings to other datasets that have the same image type and attributes. In light of this, it would be beneficial to repeat this experiment using more DL methods, such as the UNET model or various CNN model types, using additional publicly or privately available datasets in order to corroborate or refute the findings of the study.

5.3 Construct Validity

To assess the dependability of the classifier, this study concentrated on the accuracy and other performance measures (precision, recall and F1-Score). The fact that these metrics are often used to assess categorization performance was the primary factor in the selection of these performance criteria. In order to avoid favoring one performance criterion over another, the results were also obtained using the SK test and Borda count voting technique with equal weights utilizing the four performance criteria.

6 CONCLUSIONS AND FUTURE WORK

This paper presented and discussed the process and the results of an empirical evaluation study of eight end-to-end CNN architectures, notably VGG19, VGG16, ResNet50, Inception V3, DenseNet201, MobileNet V2, Xception, and Inception ResNet V2 for the binary classification of BCS mammogram images retrieved from the datasets CBIS-DDSM and DDSM. The empirical testing was based on four performance metrics (accuracy, precision, recall, and f1-score), as well as a statistical testing which consisted of Scott-Knott test and Borda count voting system. This study's findings can be summarized in two main points:

(RQ1) What is the overall performance of DL techniques in BCS binary classification?

The accuracy percentages of the eight DL techniques were all satisfactory and the majority exceeded 80%. The highest accuracies were scored by DenseNet201, MobileNet V2, and Inception Resnet V2, respectively. While ResNet50 was, compared to the other models, underperforming.

(RQ2): Are there any DL techniques that noticeably outperformed the others?

SK test resulted in only one cluster, so further testing was done to ensure that the best model is chosen. So, by ranking first in the Borda count voting system, DenseNet201 is the best performing DL architecture out of the eight trained DL models. Thus, it is a highly recommended model to serve as base for a DL computer assistance program to aid in the process of BCS.

This work came as a part of a project that puts efforts in building a comprehensive tool of computer aided diagnosis (CAD) to help improve the process of breast cancer screening in terms of radiologists' assessment as well as reducing unnecessary steps of the process to eventually result in an early diagnosis of the disease.

As future work, we aim to experiment with different ensemble techniques, in particular the techniques that use bagging or boosting, and use a variety of pre-processing steps aimed for feature selection on breast cancer screening data. Moreover, we plan to apply and evaluating different DL architectures over tabular breast cancer screening data as well.

ACKNOWLEDGMENTS

This work was conducted under the research project "Machine Learning based Breast Cancer Diagnosis and Treatment," 2020-2023. The authors would like to thank the Moroccan Ministry of Higher Education and Scientific Research, Digital Development Agency (ADD), CNRST, and UM6P for their support.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Research, G. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. https://arxiv.org/abs/1603.04467v2

Aboutalib, S. S., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., & Wu, S. (2018). Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clinical Cancer Research*, *24*(23), 5902–5909. https://doi.org/10.1158/1078-0432. CCR-18-1115

Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., & Martí, R. (2019). Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, *6*(3), 031409. https://doi.org/10.1117/1.JMI.6.3.031409

Azzeh, M., Nassif, A. B., & Minku, L. L. (2017). An empirical evaluation of ensemble adjustment methods for analogy-based effort estimation. *Journal of Systems and Software*, *103*, 36–52. https://doi.org/10.1016/j.jss.2015.01.028

Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, *157*, 19–30. https://doi.org/10.1016/J.CMPB.2018.01.011

Coleman, C. (2017). Early Detection and Screening for Breast Cancer. *Seminars in Oncology Nursing*, *33*(2), 141–155. https://doi.org/10.1016/J.SONCN.2017.02.009

Eric A. Scuccimarra. (2018). *DDSM Mammography*. Kaggle.

Fei-Fei, L., Deng, J., & Li, K. (2010). ImageNet: Constructing a large-scale image database. *Journal of Vision*, *9*(8), 1037–1037. https://doi.org/10.1167/9.8.1037

García-Lapresta, J. L., & Martínez-Panero, M. (2002). Borda Count Versus Approval Voting: A Fuzzy Approach. *Public Choice 2002 112:1*, *112*(1), 167–184. https://doi.org/10.1023/A:1015609200117

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Idri, A., & Abnane, I. (2017). Fuzzy Analogy Based Effort Estimation: An Empirical Comparative Study. *IEEE CIT 2017 - 17th IEEE International Conference on Computer and Information Technology*, 114–121. https://doi.org/10.1109/CIT.2017.29

Idri, A., Abnane, I., & Abran, A. (2018). Evaluating Pred(p) and standardized accuracy criteria in software development effort estimation. *Journal of Software: Evolution and Process*, *30*(4), e1925. https://doi.org/10.1002/SMR.1925

Idri, A., Hosni, M., & Abran, A. (2016). Improved estimation of software development effort using Classical and Fuzzy Analogy ensembles. *Applied Soft Computing*, *49*, 990–1019. https://doi.org/10.1016/J.ASOC.2016.08.012

Jelihovschi, E., Faria, J. C., & Allaman, I. B. (2014). ScottKnott: A Package for Performing the Scott-Knott Clustering Algorithm in R. *TEMA (São Carlos)*, *15*(1), 003. https://doi.org/10.5540/TEMA.2014.015.01.0003

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Lahmar, C., & Idri, A. (2022). On the value of deep learning for diagnosing diabetic retinopathy. *Health and Technology*, *12*(1), 89–105. https://doi.org/10.1007/S12553-021-00606-X/FIGURES/11

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data 2017 4:1*, *4*(1), 1–9. https://doi.org/10.1038/sdata.2017.177

McAuliffe, M. J., Lalonde, F. M., McGarry, D., Gandler, W., Csaky, K., & Trus, B. L. (2001). Medical image processing, analysis & visualization in clinical research. *Proceedings of the IEEE Symposium on Computer-Based Medical Systems*, 381–388. https://doi.org/10.1109/CBMS.2001.941749

Ottoni, A. L. C., Nepomuceno, E. G., de Oliveira, M. S., & de Oliveira, D. C. R. (2019). Tuning of reinforcement learning parameters applied to SOP using the Scott–Knott method. *Soft Computing*, *24*(6), 4441–4453. https://doi.org/10.1007/S00500-019-04206-W

Pedregosa, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, A., & Duchesnay EDOUARD DUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.5555/1953048

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, *39*(3), 355–368. https://doi.org/10.1016/S0734-189X(87)80186-X

Reback, J., jbrockmendel, McKinney, W., Bossche, J. van den, Augspurger, T., Cloud, P., Hawkins, S., Roeschke, M., gfyoung, Sinhrks, Klein, A., Hoefler, P., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Darbyshire, J., Garcia, M., … Seabold, S. (2022). *pandas-dev/pandas: Pandas 1.4.1*. https://doi.org/10.5281/ZENODO.6053272

Saranyaraj, D., Manikandan, M., & Maheswari, S. (2020). A deep convolutional neural network for the early detection of breast carcinoma with respect to hyper-parameter tuning. *Multimedia Tools and Applications*, *79*(15–16), 11013–11038. https://doi.org/10.1007/S11042-018-6560-X/TABLES/12

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-48995-4

The American Cancer Society medical and editorial content team. (2022, January 12). *Key Statistics for Breast Cancer*.

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

Worsley, K. J. (1977). A Non-Parametric Extension of a Cluster Analysis Method by Scott and Knott. *Biometrics*, *33*(3), 532. https://doi.org/10.2307/2529369

Zerouaoui, H., & Idri, A. (2022). Deep hybrid architectures for binary classification of medical breast cancer images. *Biomedical Signal Processing and Control*, *71*, 103226. https://doi.org/10.1016/J.BSPC.2021.103226

Zerouaoui, H., Idri, A., Nakach, F. Z., & Hadri, R. el. (2021). *Breast Fine Needle Cytological Classification Using Deep Hybrid Architectures*. 186–202. https://doi.org/10.1007/978-3-030-86960-1_14

Zhang, Z. (2019). Improved Adam Optimizer for Deep Neural Networks. *2018 IEEE/ACM 26th International Symposium on Quality of Service, IWQoS 2018*. https://doi.org/10.1109/IWQOS.2018.8624183

Zizaan, A., & Idri, A. (2022). *Systematic Literature review of Machine Learning based Breast Cancer Screening*.