

Breast Fine Needle Cytological Classification Using Deep Hybrid Architectures

Hasnae Zerouaoui[2], Ali Idri[1,2(✉)], Fatima Zahrae Nakach[1], and Ranya El Hadri[1]

[1] Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat, Morocco
ali.idri@um5.ac.ma

[2] Modeling, Simulation and Data Analysis, Mohammed VI Polytechnic University, Benguerir, Morocco

Abstract. Diagnosis of breast cancer in the early stages allows to significantly decrease the mortality rate by allowing to choose the adequate treatment. This paper develops and evaluates twenty-eight hybrid architectures combining seven recent deep learning techniques for feature extraction (DenseNet 201, Inception V3, Inception ReseNet V2, MobileNet V2, ResNet 50, VGG16 and VGG19), and four classifiers (MLP, SVM, DT and KNN) for binary classification of breast cytological images over the FNAC dataset. To evaluate the designed architectures, we used: (1) four classification performance criterias (accuracy, precision, recall and F1-score), (1) Scott Knott (SK) statistical test to cluster the developed architectures and identify the best cluster of the outperforming architectures, and (2) Borda Count voting method to rank the best performing architectures. Results showed the potential of combining deep learning techniques for feature extraction and classical classifiers to classify breast cancer in malignant and benign tumors. The hybrid architectures using MLP classifier and DenseNet 201 for feature extraction were the top performing architectures with a higher accuracy value reaching 99% over the FNAC dataset. As results, the findings of this study recommend the use of the hybrid architectures using DenseNet 201 for the feature extraction of the breast cancer cytological images since it gave the best results for the FNAC data images, especially if combined with the MLP classifier.

Keywords: Computer-aided diagnosis · Breast cancer · Classification · Deep convolutional neural networks · Image processing · Histological images · Hybrid architecture

1 Introduction

Breast cancer (BC) is the most commonly diagnosed cancer for women worldwide with an estimate of 2.3 million new cases in 2020 [1, 2]. It remains a global challenge, since it is the fifth leading cause of mortality worldwide, and the second cause of death for women globally [2, 3]. As the number of patients affected by the BC disease increases, it turns out to be hard for radiologists, pathologists and oncologists to accurately deal with the diagnosis process in the constrained accessible time [4]. Medical images analysis

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12950, pp. 186–202, 2021.
https://doi.org/10.1007/978-3-030-86960-1_14

is one of the most promising research areas, and it provides facilities for diagnosis and decision making of several diseases such as breast cancer. Lately, more attention are paid to imaging modalities and deep learning (DL) in BC [5–7]. Therefore, interpretation of these images requires expertise and consequently several techniques have been developed and evaluated to improve and help oncologist's diagnosis. In general, deep learning techniques showed better performance in breast cancer detection by extracting the most important features of medical images, and provided high accurate features compared with classical features extraction techniques of the image processing process [6–9]. In the other hand, classical machine learning techniques gives accurate results for BC classification, are less time consuming and need fewer parameter tunning compared to the ones of the DL techniques [6, 10]. For instance, many researchers designed hybrid architectures where they combined the strengths of DL techniques for feature extraction and classical machine learning for classification [11–13]. For reference, the study [10] showed that the use of hybrid architectures combining deep CNN architectures and recurrent neural networks (RNNs) gave better results compared to CNN and deep CNN; indeed, the proposed solution combined GoogleNet and AlexNet for features extraction and LSTM (long short-term memory) for classification over the BreakHis dataset, and it achieved an accuracy of 90.5%. In [14], the authors designed a hybrid architecture using AlexNet and VGG16 DL techniques for feature extraction over the BreakHis dataset; once the features extracted, they concatenated feature vectors and apply the SVM classifier for a binary classification, the best accuracy results achieved using AlexNet and VGG16 for feature extraction and SVM for classification are: 84.87%, 89.21%, 8.65% and 86.75% for the MF 40X, 100X, 200X and 400X respectively.

This paper develops and evaluates twenty-eight hybrid architectures using four classifiers (MLP, SVM, DT and KNN) and seven of the most recent DL techniques as feature extractor (DenseNet 201, Inception V3, Inception ReseNet V2, MobileNet V2, ResNet 50, VGG16 and VGG19) for a binary BC classification since [7]: (1) they are frequently used for diagnosing breast cancer patients since 2016, and (2) they proved their strength for FE compared to the classical ones; the designed architectures are tested the FNAC dataset. We evaluate the performance of the designed hybrid architectures in terms of accuracy, precision, recall and F1-score. To the best of our knowledge, this study is the first to evaluate and compare twenty-eight hybrid architectures using seven DL techniques for feature extraction and four classifiers and uses the Scott Knott (SK) statistical test and Borda count voting method in BC image processing and classification. Note that the SK test has been widely used to comparing, clustering and ranking multiple machine learning models for parameters tunning [18, 19] in different fields such as software engineering [7, 8] and breast cancer [22]. Hence, we use the SK test since: (1) its high performance compared to other statistical tests such as Jollife [21], Calinski and Corsten [24], and Cox and Spjotvoll [25], and (2) its ability to cluster the best non-overlapping groups of machine learning techniques. Furthermore, we use the Borda Count voting method [14, 15] to rank the best SK hybrid architectures. The present study discusses three research questions (RQs):

- (RQ1): What is the overall performance of the hybrid architectures in BC binary classification?

- (RQ2): Is there any deep learning technique for feature extraction which distinctly outperformed the others when used in the hybrid architectures constructed for each classifier?
- (RQ3): Is there any hybrid architectures which distinctly outperformed the others regardless the feature extractor technique and the classifier used?

The main contributions of this empirical study are the following:

1. Designing twenty-eight hybrid architectures using four classifiers: MLP, SVM, DT and KNN, and seven DL architectures for feature extraction (FE): VGG16, VGG19, DenseNet201, Inception ResNet V2, Inception V3, ResNet 50 and MobileNet V2 in BC binary classification.
2. Evaluating the twenty-eight hybrid architectures over the FNAC dataset.
3. Comparing the performances of the twenty-eight hybrid architectures using SK clustering test and Borda Count voting method.

The rest of this paper is organized as follow. Section 2 presents the data preparation which includes data acquisition, image pre-processing and data augmentation. Section 3 presents the empirical methodology followed in this research. Section 4 reports and discusses the empirical results. Section 5 presents the threats of validity of the study. Section 6 outlines conclusions and future works.

2 Data Preparation

This section presents the data preparation process we followed for the FNAC dataset, which consists of data acquisition, data pre-processing and data augmentation.

2.1 Data Acquisition

In this subsection, we present the publicly available Fine Needle Aspiration Cytology (FNAC) dataset used in all the empirical evaluations. The images of the FNAC dataset were captured by using Leica ICC50 HD microscope using 400 resolution and 24 bits color depth and with 5 megapixels camera associated with the microscope [33]. Digitized images captured were then reviewed by experienced certified cyto-pathologists and selected a total of 212 images (113 Malignant and 99 Benign). The database can be downloaded from the link in [33].

2.2 Data Processing

In the subsection, we present the image processing we used in this study for the FNAC dataset. The first and most important step of the data preparation is to preprocess input images using different pre-processing techniques [34]. In this study, we used intensity normalization and Contrast Limited Adaptive Histogram Equalization (CLAHE). Intensity normalization is an interesting pre-processing step in image processing applications

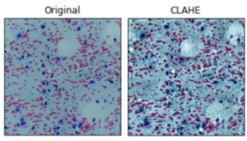

Fig. 1. (a). Original image, (c). CLAHE

and is highly used to preprocess medical images [6, 13, 35]. As shown in Fig. 1, we normalized input images to the standard normal distribution using min-max normalization of Eq. 1, and we improved the contrast using the CLAHE technique [32, 36].

$$X_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

2.3 Data Augmentation

Data augmentation was used for the training process after dataset pre-processing and has the goal to avoid the risk of overfitting [37]. Moreover, the strategies we used include geometric transforms such as rescaling, rotations, shifts, shears, zooms and flips as shown in Table 1.

Table 1. Data augmentation used

Argument	Parameter value	Description
Rescale	1/255.0	Scale images from integers 0–255 to floats 0–1
Rotation range	90	Degree range of the random rotations
Vertical shift range	0.2	The parameter value of horizontal and vertical shifts (20%) is a fraction of the given dimension
Horizontal shift range	0.2	
Shear range	0.2	Controls the angle in counter clockwise direction as radians in which our image will allowed to be sheared
Zoom range	0.2	Allows the image to be "zoomed out" or "zoomed in"
Horizontal flip	True	Controls when a given input is allowed to be flipped horizontally during the training process
Fill mode	Nearest	This is the default option where the closest pixel value is chosen and repeated for all the empty values

3 Empirical Design

This section presents the empirical design we followed in this study including: (1) performance criteria we used to evaluate the models and 10-fold cross validation to

evaluate the hybrid architectures, (3) Skott Knott statistical test we used to cluster the hybrid architectures according to their accuracy values, (4) Borda Count method we used to rank the hybrid architectures of the best SK cluster according to accuracy, precision, recall, and F1-core, (5) experimental process we followed to carry out all the empirical evaluations, (6) the abbreviations we used to shorten the names of the hybrid architectures developed, and (7) the experiment configuration of the classifiers and DL feature extractors.

3.1 Performance Measures

To evaluate the performance of the Hybrid techniques, we used the followed performance measures [6]: accuracy, precision, recall and F1-score, defined in the Eq. (2), (3), (4) and (5).

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2 \times \frac{Recall \times precision}{Recall + precision} \tag{5}$$

As for the evaluation, we used 10-fold cross validation since it is easy to understand and gives better performance results compared to other evaluation techniques [71–73]. We reported the average of the performance of each of the 10 folds.

3.2 Statistical Tests

Scott Knott: The Scott-Knott (SK) algorithm is an exploratory clustering algorithm usually used in the analysis of variance (ANOVA) context. It was proposed by Scott and Knott in 1974 [38] to find distinction overlapping groups based on the multiple comparisons of treatment means. The SK algorithm is the most frequently used hierarchical clustering algorithm due to its simplicity and robustness [27–29]. In this study the SK test was performed based on the accuracy performance measure.

Borda Count: Borda count is a voting method for single winner election methods. In this technique, points are given to candidates based on their ranking; 1 point for last choice, 2 points for second-to-last choice, and so on until you are at the top. The point values for all ranks are totaled, and the candidate with the largest point total is the winner [26]. In the present study, we used Borda count technique to find the best performing hybrid architectures from the four performance measures with equal weights. This strategy was adopted to make sure that we do not favor a measure. This voting technique is widely used in SDEE, indeed [20, 30, 31], deduced that the conclusion on which a technique is the most performing depends on the chosen performance measures

that may lead to contradictory results: for instance each technique may have different voting according to different performance measures. The borda count was used in this study based on four performance metrics: accuracy, precision, recall and F1-score.

3.3 Experimental Process:

Figure 2 shows the methodology we followed to carry out the empirical evaluations of this experiment. It consists of five steps described hereafters. Note that similar methodologies were used in [8, 9, 35, 36, 37]. The evaluation process involves:

1. Assess the accuracy of each variant of the four classifiers (MLP, SVM, DT, KNN) using the seven deep learning architectures as feature extractors (VGG16, VGG19, DensNet201, MobileNet_V2, ResNet50, Inception_V3, Inception_ResNet_V2) for FNAC dataset.
2. Clustering the four classifiers (MLP, SVM, DT, KNN) using Scott Knott test based on accuracy for both datasets and select the best hybrid architecture of the best SK cluster (have the best accuracy and statistically indifferent).
3. Rank the classifiers of the best SK cluster using Borda count voting system based on the four performance measures (accuracy, precision, recall and F1-score) and select the top hybrid architecture for each dataset.
4. Select the best ranked hybrid architectures for each classifier from step three and apply SK test based on accuracy.
5. Rank the best cluster of step four using Borda count based on the four performance metrics: accuracy, precision, recall and f1-score.

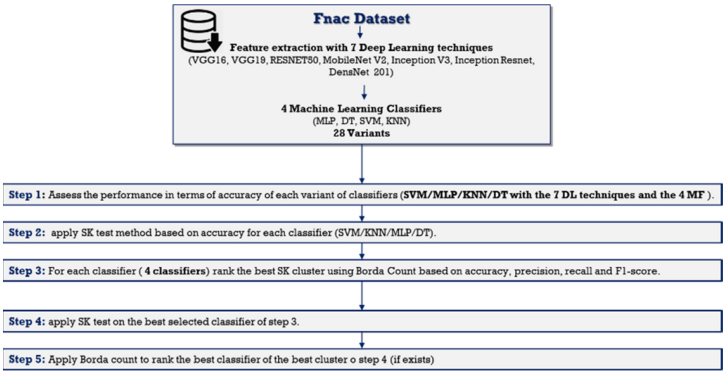

Fig. 2. Experimental process

3.4 Abbreviation

In order to shorten the names of the hybrid architectures, we use the following naming rules in the rest of this paper: *Classifier Deep Learning Architecture*

We shorten the name of the classifiers as follow: M for MLP, S for SVM, D for DT and K for KNN. And for the deep learning architectures we choose: V16 for VGG16, V19 for VGG19, IN for Inception V3, INR for InceptionResNet V2, Res for ResNet 50, MOB for MobilNet V2 and DEN for DensNet 201.

3.5 Experiment Configuration

To design the hybrid architectures using four classifiers (MLP, SVM, DT and KNN) and seven DL techniques for FE (VGG16, VGG19, Inception V3, DenseNet 201, ResNet 50, Inception ResNet V2 and MobileNet V2) for both datasets we used the following experiment configurations:

- All the images of the FNAC dataset were resized to 224 × 224 pixels except those of Inception_V3 and Inception_ResNet_V2 models that were resized to 299 × 299 since it is the default input size in their architectures.
- To extract the features of the input images we used the transfer learning [38] technique for FE where we downloaded the seven DL techniques pre-trained in the ImageNet dataset from the Keras library, we froze the top layers and generated a new dataset containing the features of FNAC dataset that will be the input of the four classifiers (MLP, SVM, DT and KNN).
- For binary classification we used the default configuration of the four classifiers (MLP, SVM, DT and KNN) given by the scikit-learn library of python programming language.

4 Results and Discussion

This section presents and discusses the results of the empirical evaluations of the hybrid architectures over FNAC dataset. We firstly evaluate the performances in terms of accuracy of the hybrid architectures of each classier (RQ1). Thereafter, we evaluate the impacts of the seven DL feature extraction techniques on the performances of the four classifiers in order to identify which one of them are positively influencing the classification performance (RQ2). Lastly, we compare the best hybrid architectures of the four classifiers in order to identify which hybrid architecture performed better over the FNAC dataset.

4.1 (RQ1): What is the Overall Accuracy Performance of the Hybrid Architectures in a BC Binary Classification?

To evaluate and compare hybrid architectures, we assess the accuracy values for each classifier. Note that the empirical evaluations of the hybrid architectures were performed using a computer with Processor: Intel (R) core (TM) i7-7500 CPU @ 2.90 GHZ and 12 Go in RAM running on a Microsoft Windows 10 Professional (64-bit). Python 3.7.4 was used with the two DL frameworks Keras and Tensorflow as deep learning backend, and the scikit-learn framework for machine learning. R 3.4.4 was used for statistical evaluation.

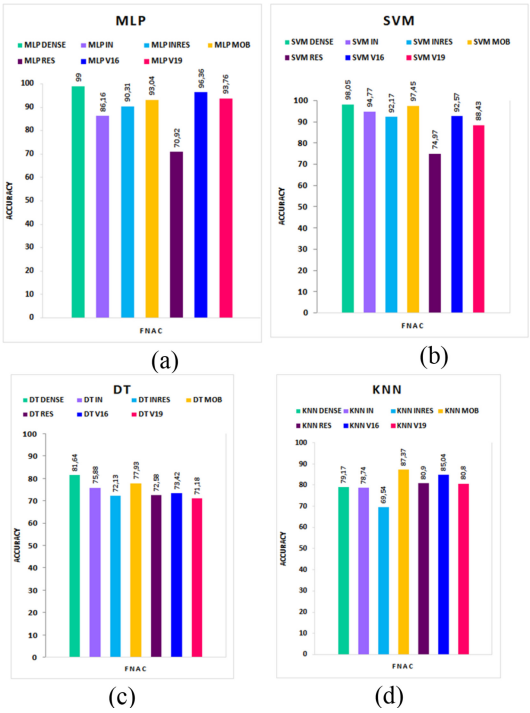

Fig. 3. Accuracy values for the twenty-eight hybrid architectures of the FNAC dataset

This subsection presents and discusses the performances of the hybrid architectures of each classifier over the FNAC dataset. Table 2 summarizes the testing accuracy values. Figures 3 shows the mean accuracy values of the 4 classifiers depending on the DL feature extractor used.

We observe that:

- For the hybrid architectures using MLP classifier the best accuracy value was obtained when using the DenseNet 201 for FE and it reached 99% and the worst accuracy value when using ResNet 50 for FE was 70.92%.
- For the hybrid architectures using the SVM classifier, the best accuracy value was obtained when using DenseNet 201 for FE and reached 98.05% and the worst accuracy value when using ResNet 50 for FE was 74.97%.
- For the hybrid architectures using the DT classifier, the best accuracy value was obtained when using DenseNet 201 for FE and reached 81.64% and the worst accuracy value when using VGG 19 for FE was 71.18%.
- For the hybrid architectures using the KNN classifier the best accuracy value was obtained when using MobileNet V2 for FE and reached 87.37% and the worst accuracy value when using Inception Inception ResNet V2 for FE was 69.54%.

4.2 (RQ2): Is There Any Deep Learning Technique for Feature Extraction Which Distinctly Outperformed the Others When Used in a Hybrid Architecture?

This section aims at evaluating the impacts of the seven DL feature extraction techniques on the performances of the four classifiers in order to identify (if exist) the feature extraction techniques that are positively impacting the classification performance. To this aim, (1) we used the SK test based on the accuracy values of the hybrid architectures of each classifier in order to cluster the techniques that have the same predictive capabilities regardless the feature extractor used, and (2) we used the Borda count method based on the four-performance metrics: accuracy, precision, recall and F1-score to rank the hybrid architectures of each classifier belonging to the best SK clusters over both the FNAC dataset. Tables 2 shows the values of the four performance measures of all the hybrid architectures. Figure 4 shows the results of SK test over the FNAC dataset. We observe that:

- For MLP, we obtained 2 clusters which implies that regardless the DL feature extractor technique used, we obtained statistically similar results in terms of accuracy. The best SK cluster contains 6 hybrid architectures using MLP with DenseNet 201, VGG16, VGG19, MobileNet V2, Inception Resnet V2, or Inception V3 for feature extraction. The last cluster contains MLP with ResNet 50.
- For SVM, we obtained 5 clusters which implies that the SVM classifier is impacted by the DL feature extractor used. The best SK cluster contains 2 hybrid architectures: SVM with DenseNet 201 or MobileNet V2. The second SK cluster contains SVM with Inception V3 as feature extractor. The third cluster, we find the architectures constructed by SVM with the DL techniques VGG16 and Inception Resnet V2. The fourth cluster includes the hybrid architecture SVM with VGG19. The last cluster contains SVM with ResNet 50 as feature extractor.
- For DT, the best cluster contains one architecture: DT with DenseNet 201. The second cluster contains the architecture combining DT and MobileNet V2. In the third cluster we have the architecture combining DT and Inception V3. For the last cluster we have the hybrid architectures constructed by the DT classifier and VGG16, ResNet 50, Inception Resnet V2 and VGG19 for feature extraction.

For KNN, 3 clusters where identified: the best SK cluster containing 2 hybrid architectures: KNN with MobileNet V2 or VGG16. The second SK cluster contains the architectures KNN with ResNet 50, VGG 19, DenseNet 201 or Inception V3. The last cluster contains KNN with Inception Resnet V2 as feature extractor.

Table 3 shows the Borda count ranks of the best SK clusters using the four classifiers over the FNAC dataset. As can be seen, the hybrid architectures designed with DensNet 201 were ranked first regardless the classifier, except for the KNN classifier. The second feature extractor is MobileNet V2 since it was ranked first with KNN, second using SVM and forth with MLP. As for the feature extractor VGG16, it was ranked second for both MLP and KNN. For the feature extractors VGG19, Inception V3 and InceptionResNet V2 they only belong to the best cluster with the MLP classifier and were ranked third, fifth and sixth respectively. Finally ReseNet 50 feature extractor underperformed compared to the others. In order to rank the impacts of each feature extractor on the classification

Table 2. Accuracy, precision, recall and F1 score values of the hybrid architectures over the FNAC dataset

Classifier	Hybrid architectures	Accuracy	Precision	Recall	F1-score
MLP	MDEN	0.99	0.9956	0.9849	0.9902
	MIN	0.8616	0.8166	0.8528	0.8259
	MINR	0.9032	0.8675	0.8355	0.8503
	MMOB	0.9305	0.8803	0.8714	0.8756
	MRES	0.7092	0.6276	0.644	0.6039
	MV16	0.9636	0.9666	0.9585	0.9623
	MV19	0.9377	0.953	0.9172	0.9343
SVM	SDEN	0.9805	0.9886	0.9729	0.9806
	SIN	0.9478	0.9563	0.936	0.9458
	SINR	0.9217	0.9489	0.8877	0.9168
	SMOB	0.9746	0.9829	0.965	0.9738
	SRES	0.7498	0.8494	0.5898	0.6938
	SV16	0.9257	0.9481	0.8982	0.9221
	SV19	0.8843	0.9304	0.8263	0.8739
DT	DDEN	0.8164	0.8379	0.8139	0.8247
	DIN	0.7589	0.7517	0.7606	0.7558
	DINR	0.7213	0.7235	0.6979	0.7086
	DMOB	0.7793	0.7692	0.7857	0.7761
	DRES	0.7258	0.7186	0.7225	0.7186
	DV16	0.7343	0.7376	0.7051	0.7201
	DV19	0.7118	0.7076	0.696	0.7011
KNN	KDEN	0.7918	0.9148	0.6684	0.7713
	KIN	0.7875	0.9791	0.58	0.7273
	KINR	0.6954	0.8941	0.4254	0.5746
	KMOB	0.8737	0.9925	0.7475	0.852
	KRES	0.8091	0.8756	0.7108	0.7826
	KV16	0.8504	0.9115	0.769	0.8332
	KV19	0.8081	0.88	0.7017	0.78

performance regardless the classifier, we count the number of occurrences of each feature extractor on the best SK clusters. In case of a tie, we refer to the borda count ranking. As can be seen in Table 4, the best performing feature extractor is DenseNet 201, since it appears 3 times in the best SK cluster and was ranked 3 times first. The following feature extractor is MobileNet V2, it appears 3 times in the best SK cluster and was

ranked first, second and forth. As for the remaining DL technique VGG19, Inception V3 and Inception ReseNet V2 they appeared 1 time in the best SK cluster. Finally, Resnet 50 feature extractor underperform compared the others, and never appear in the best SK cluster. As for the sensitivity of the classifiers to the feature extraction technique used, we notice that MLP is less sensitive compared to the others. In fact, the numbers of SK clusters of MLP are 2, while those of KNN are 3, for DT 4 and finally 5 for SVM.

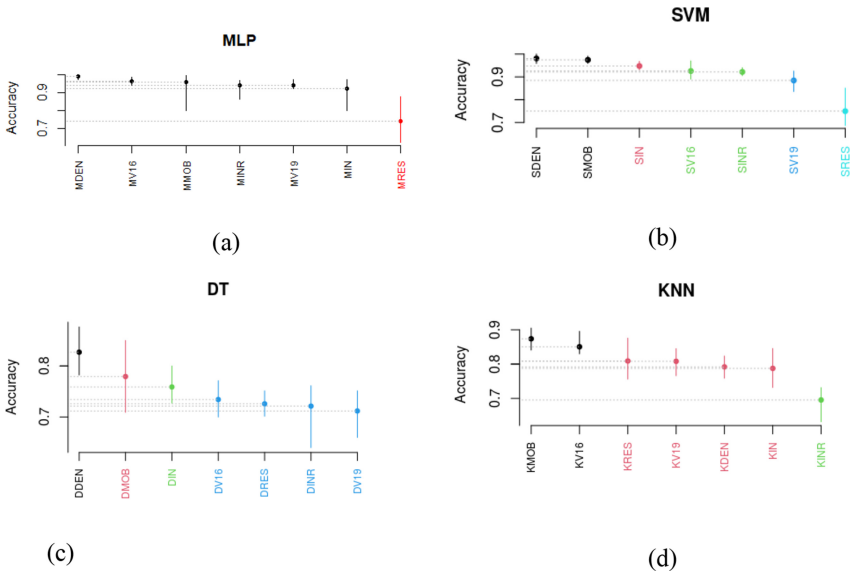

Fig. 4. Results of SK test for the hybrid architecture using the four classifiers over the FNAC dataset

Table 3. Borda count ranking fort the hybrid architectures of the best SK cluster

Classifier ranking	MLP	SVM	DT	KNN
1	MDEN	SDEN	DDEN	KMOB
2	MV16	SMOB		KV16
3	MV19			
4	MMOB			
5	MINR			
6	MIN			

Table 4. Appearance in the best SK cluster and borda count ranking

DL for feature extraction	Appearance in the best SK cluster	Borda count ranking
DenseNet 201	3	3 times first rank
MobileNet V2	3	1 time first rank 1 time second rank 1 time forth rank
VGG16	2	2 times second rank
VGG19	1	Third rank
ReseNet 50	–	–
Inception V3	1	Sixth rank
Inception ReseNet V2	1	Fifth rank

4.3 (RQ3): Is There Any Hybrid Architectures Which Distinctly Outperformed the Others Regardless the Feature Extractor and the Classifier Used?

This Section uses the SK test based on accuracy to evaluate the predictive capabilities of the best hybrid architectures of the four classifiers over each dataset (Step 4 of the empirical design). Thereafter, we discuss the ranking results of the borda count voting method based on accuracy, precision, recall and F1-score on the best SK clusters to identify the best hybrid architecture regardless the classifier over the FNAC dataset. Figure 5 shows the results of SK test over the best ranked hybrid architecture of the FNAC dataset regardless the classifier used. We obtained three clusters. The best SK cluster contains the two hybrid architectures MLP and SVM with DenseNet 201. The second SK cluster contains the KNN classifier with MobileNet V2. And the last cluster contains DT with DenseNet.

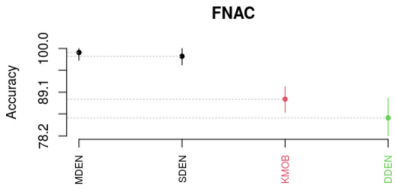

Fig. 5. Results of SK test for the best ranked hybrid architectures over the FNAC dataset

Using the Borda count voting methods on the best SK cluster, we observe from Table 5 that the hybrid architectures MLP with DenseNet 201 was ranked first, and the hybrid architecture SVM with DenseNet 201 was ranked second. As regard we conclude that MDEN hybrid architecture outperforms the others.

Table 5. Borda count ranking for the best SK cluster oft he best ranked hybrid architectures constructed over the FNAC dataset

Rank	FNAC dataset hybrid architectures
1	MDEN
2	SDEN

5 Threats of Validity

This section presents the threats to this paper's validity with respect to external and internal validity.

Internal Validity: This paper used the 10-fold cross validation method for evaluation. The main reason to use the 10-fold cross validation is that cross validation gives more stable estimations [39]. Another internal threat for this experiment is the use of transfer learning for feature extraction while using the DL techniques to extract the features of the images of the FNAC dataset.

External Validity: This study used one dataset which contain cytological images; so, we cannot generalize the obtained results for all the datasets with the same type of images and the same features. However, it will be a good benefit to redo this study using different DL techniques with another publicly or private datasets in order to confirm or refute the findings of this study.

Construct Validity: For the reliability of the classifier performances obtained, this study focused on the accuracy and other four performance criteria (accuracy, precision, recall and F1-score). The main reasons behind the choice of these performance criteria are: (1) most of the studies used them to measure the classification performance [7], and (2) the type of the data is balanced. Moreover, the conclusion was drawn by using the SK test and Borda count voting system with equal weights using these four performance criteria (accuracy, precision, recall and F1-Score). This strategy was adopted to make sure that we do not favor a particular performance criterion than another.

6 Conclusion and Future Work

The present paper presented and discussed the results of an empirical comparative study of 28 hybrid architectures using four classifiers (MLP, SVM, DT, and KNN) and seven DL techniques for feature extraction (DensNet 201, MobileNet V2, ReseNet 50, Inception V3, Inception ReseNet V2, VGG16 and VGG19), for BC Imaging classification. All the empirical evaluations used four performance criteria (accuracy, precision, recall and F1-score), SK statistical test, and Borda Count to assess and rank these 28 hybrid architecturs over the FNAC database. The main findings of this study are:

(RQ1): What is the Overall Performance of the Hybrid Architectures in BC Binary Classification?

The accuracy results of the constructed hybrid architectures were highly influenced by the DL techniques used for feature extraction. As regard, we observed that the use of DenseNet 201 for FE in the design of the hybrid methods almost gave the best results and outperformed the others regardless the classifier used in this experiment. However, ResNet 50 for FE underperformed the others regardless the classifiers used in the hybrid architectures. In respect we recommend the use of DenseNet DL technique for feature extraction for the design of hybrid architectures regardless the classifier used.

(RQ2): Is There Any Deep Learning Technique for Feature Extraction Which Distinctly Outperformed the Others When Used in the Hybrid Architectures Constructed for Each Classifier?
The hybrid architectures with DenseNet 201 for feature extraction regardless the classifier used gave the best results since they always belong to the best SK cluster for the FNAC dataset and was ranked first. Furthermore, the architectures using MobileNet V2 belong three times to the best SK followed by the architectures using VGG16 that belongs two times to the best SK cluster. Regardless the classifier, the architectures using Resnet 50 for FE never belongs to the best SK cluster and is underperforming compared to the others. As result, we highly recommend the use of DenseNet 201 to design the hybrid architectures since it gives the best results.

(RQ3): Is There Any Hybrid Architectures Which Distinctly Outperforms the Others Regardless the Feature Extractor Technique and the Classifier Used?
Step four and five of the empirical design have allowed us to distinguish the most outperforming constructed hybrid architecture regardless the classifier and for FNAC datasets. In this sense the MDEN hybrid architecture constructed using the MLP classifier and DenseNet 201 DL technique for feature extraction gave the best results since it belongs to the best cluster and was ranked first using the borda count method. The accuracy values achieved by the MDEN hybrid architecture are: 99% for the FNAC dataset.

In the future work, we are planning to investigate other public dataset to validate or refute the finding of this study. Another perspective for our research is to build homogenous and heterogeneous ensembles using different DL techniques for feature extraction for breast cancer imaging classification.

Acknowledgement. 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.

This study was funded by Mohammed VI polytechnic university at Ben Guerir Morocco.

Abbreviation

In order to shorten the names of the hybrid architectures We use the following naming rules in the rest of this paper: Classifier DeepLearningArchitecture

References

1. Metelko, Z., et al.: Pergamon the world health organization quality of life (WHOQOL): position paper from WHO. Soc. Sci. Med. **41**(10), 1403–1409 (1995)
2. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**(3), 1–41 (2021)
3. Bish, A., Ramirez, A., Burgess, C., Hunter, M.: Understanding why women delay in seeking help for breast cancer symptoms. J. Psychosom. Res. **58**(4), 321–326 (2005)
4. Zhang, G., Wang, W., Moon, J., Pack, J.K., Jeon, S.I.: A review of breast tissue classification in mammograms. In: Proceedings of the 2011 ACM Symposium on Research in Applied Computation RACS 2011, pp. 232–237 (2011)
5. Mendelson, E.B.: Imaging: potentials and limitations. Am. J. Roentgenol. **212**(2), 293–299 (2019). https://doi.org/10.2214/AJR.18.20532
6. Zerouaoui, H., Idri, A.: Reviewing machine learning and image processing based decision-making systems for breast cancer imaging. J. Med. Sys. **45**(1), 1–20 (2021)
7. Zerouaoui H., Idri A., El Asnaoui K.: Machine Learning and Image Processing for Breast Cancer: A Systematic Map. In: Rocha Á., Adeli H., Reis L., Costanzo S., Orovic I., Moreira F. (eds.) Trends and Innovations in Information Systems and Technologies. WorldCIST 2020. Advances in Intelligent Systems and Computing, vol. 1161. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45697-9_5
8. Idri, A., Chlioui, I., El Ouassif, B.: A systematic map of data analytics in breast cancer. In: International Conference Proceedings Series (2018)
9. Ouassif, E., Idri, A., Hosni, M., Abran, A.: Classification techniques in breast cancer diagnosis: a systematic literature review. Comput. Methods Biomech. Biomed. Eng. Imag. Vis. **9**(1), 50–77 (2000)
10. Yan, R. et al.: A hybrid convolutional and recurrent deep neural network for breast cancer pathological image classification. In: International Conference on Bioinformatics and Biomedicine (BIBM 2018), pp. 957–962 (2019)
11. Mendel, K., Li, H., Sheth, D., Giger, M.: Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography. Acad. Radiol. **26**(6), 735–743 (2019)
12. Valueva, M.V., Nagornov, N.N., Lyakhov, P.A., Valuev, G.V., Chervyakov, N.I.: ScienceDirect application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Math. Comput. Simul. **177**, 232–243 (2020)
13. Cordeiro, F.R., Santos, W.P., Silva-Filho, A.G.: A semi-supervised fuzzy GrowCut algorithm to segment and classify regions of interest of mammographic images. Expert Syst. Appl. **65**, 116–126 (2016)
14. Deniz, E., Şengür, A., Kadiroğlu, Z., Guo, Y., Bajaj, V., Budak, Ü.: Transfer learning based histopathologic image classification for breast cancer detection. Heal. Inf. Sci. Syst. **6**(1), 18 (2018)
15. Abdar, M., Makarenkov, V.: CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement **146**, 557–570 (2019)
16. Ottoni, A.L.C., Nepomuceno, E.G., de Oliveira, M.S., de Oliveira, D.C.R.: Tuning of reinforcement learning parameters applied to SOP using the Scott-Knott method. Soft Comput. **24**(6), 4441–4453 (2020)
17. Idri, A., Hosni, M., Abran, A.: Improved estimation of software development effort using Classical and fuzzy analogy ensembles. Appl. Soft Comput. J. **49**, 990–1019 (2016)
18. Mittas, N., Angelis, L.: Ranking and clustering software cost estimation models through a multiple comparisons algorithm. IEEE Trans. Softw. Eng. **39**(4), 537–551 (2013)

19. Mittas, N., Mamalikidis, I., Angelis, L.: A framework for comparing multiple cost estimation methods using an automated visualization toolkit. Inf. Softw. Technol. **57**(1), 310–328 (2015)
20. Idri, A., Bouchra, E., Hosni, M., Abnane, I.: Assessing the impact of parameters tuning in ensemble based breast cancer classification. Health Technol. (Berl) **10**(5), 1239–1255 (2020)
21. Jolliffe, I.T., Allen, O.B., Christie, B.R.: Comparison of variety means using cluster analysis and dendrograms. Exp. Agric. **25**(2), 259–269 (1989). https://doi.org/10.1017/S0014479700016768
22. Calinski, T., Corsten, L.C.A.: Clustering means in ANOVA by simultaneous testing. Biometrics **41**(1), 39 (1985)
23. Worsley, K.J.: Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika **73**(1), 91–104 (1986)
24. Emerson, P.: The original Borda count and partial voting. Soc. Choice Welfare **40**(2), 353–358 (2013)
25. García-Lapresta, J.L., Martínez-Panero, M.: Borda count versus approval voting: a fuzzy approach. Public Choice **112**(1), 167–184 (2002)
26. Black, D.: Partial justification of the Borda count. Public Choice **28**(1), 1–15 (1976)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015) - Conference Track Proceedings, pp. 1–14 (2015)
28. Szegedy, C., Vanhoucke, V., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (2014)
29. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (*AAAI'17*). AAAI Press, pp. 4278–4284 (2017)
30. Saikia, A.R., Bora, K., Mahanta, L.B., Das, A.K.: Comparative assessment of CNN architectures for classification of breast FNAC images. Tissue Cell **57**, 8–14 (2019)
31. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: overview, challenges and the future. In: Dey, N., Ashour, A., Borra, S. (eds.) Classification in BioApps. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65981-7_12
32. Kharel, N., Alsadoon, A., Prasad, P.W.C., Elchouemi, A.: Early diagnosis of breast cancer using contrast limited adaptive histogram equalization (CLAHE) and morphology methods. In: 2017 8th International Conference on Information and Communication Systems (ICICS 2017), pp. 120–124 (2017)
33. Makandar, A., Halalli, B.: Breast cancer image enhancement using median filter and CLAHE. Int. J. Sci. Eng. Res. **6**(4), 462–465 (2015)
34. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning (2017)
35. Azzeh, M., Nassif, A.B., Minku, L.L.: An empirical evaluation of ensemble adjustment methods for analogy-based effort estimation. J. Syst. Softw. **103**, 36–52 (2015)
36. Idri, A., Abnane, I., Abran, A.: Evaluating Pred(p) and standardized accuracy criteria in software development effort estimation. J. Softw. Evol. Process **30**(4), 1–15 (2018)
37. Idri, A., Abnane, I.: Fuzzy analogy based effort estimation: an empirical comparative study. In: IEEE CIT 2017-17th IEEE International Conference on Computer and Information Technology, no. Ml, pp. 114–121 (2017)

38. Chougrad, H., Zouaki, H., Alheyane, O.: Deep convolutional neural networks for breast cancer screening. Comput. Methods Programs Biomed. **157**, 19–30 (2018)
39. Xu, Y., Goodacre, R.: On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J. Anal. Test. **2**(3), 249–262 (2018)