

Random Forest Based Deep Hybrid Architecture for Histopathological Breast Cancer Images Classification

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

[1] Modeling, Simulation and Data Analysis, Mohammed VI Polytechnic University, Benguerir, Marrakech-Rhamna, Morocco

[2] Software Project Management Research Team, ENSIAS, Mohammed V University, Rabat-Salé-Kénitra, Rabat, Morocco

ali.idri@um5.ac.ma

Abstract. Breast cancer is the most common cancer in women worldwide. While the early diagnosis and treatment can significantly reduce the mortality rate, it is a challenging task for pathologists to accurately estimate the cancerous cells and tissues. Therefore, machine learning techniques are playing a significant role in assisting pathologists and improving the diagnosis results. This paper proposes a hybrid architecture that combines: three of the most recent deep learning techniques for feature extraction (DenseNet_201, Inception_V3, and MobileNet_V2) and random forest to classify breast cancer histological images over the BreakHis dataset with its four magnification factors: 40X, 100X, 200X and 400X. The study evaluated and compared: (1) the developed random forest models with their base learners, (2) the designed random forest models with the same architecture but with a different number of trees, (3) the decision tree classifiers with the best random forest models and (4) the best random forest models of each feature extractor. The empirical evaluations used: four classification performance criteria (accuracy, sensitivity, precision and F1-score), 5-fold cross-validation, Scott Knott statistical test, and Borda Count voting method. The best random forest model achieved an accuracy mean value of 85.88%, and was constructed using 9 trees, 200X as a magnification factor, and Inception_V3 as a feature extractor. The experimental results demonstrated that combining random forest with deep learning models is effective for the automatic classification of malignant and benign tumors using histopathological images of breast cancer.

Keywords: Breast cancer · Hybrid · Random forest · Transfer learning

1 Introduction

Breast cancer (BC) became the most common cancer globally as of 2021, accounting for 12% of all new annual cancer cases worldwide, according to the World Health Organization [1]. When BC is detected and treated early, the chances of survival are very high [2]. Computer aided detection and machine learning (ML) based intelligent automated prediction systems could improve the cancer diagnosis capability and reduce

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13376, pp. 3–18, 2022.
https://doi.org/10.1007/978-3-031-10450-3_1

the diagnosis errors [3]. Moreover, these systems can provide decision support for the doctors for an opportunity of early identification of BC [4]. Research indicates that most experienced physicians can diagnose cancer with 79% accuracy while 91% correct diagnosis is achieved using ML techniques [5].

Over the past few decades, ensemble learning has drawn the attention of computational intelligence and ML researchers [6]. This increasing interest owes to its effective role at solving a large variety of ML problems, including feature selection, pattern recognition, confidence estimation, incremental learning, surfacing in application areas such as computer vision, medical research, geosciences, and many more [7, 8]. Ensemble learning methods were originally developed to improve the prediction power and stability of the classification model by combining several learners together [9]. In fact, theoretical and empirical studies have demonstrated that an ensemble of classifiers is typically more accurate than a single classifier [10]. However, improvement over the single best classifier or even on the average of the individual accuracies is not always guaranteed [11]. This improvement of performances relies on the concept of diversity [12], that is one of the three pillars of ensemble learning [8] (the other pillars are: generating ensemble members and combining them). The practical trade-off between diversity and accuracy of the ensemble learning is still an open question in ML [13] and many methods have been proposed to generate accurate and diverse sets of models [14]. The widely-used ensemble classification methods include bagging, boosting and stacking [15].

Deep learning (DL) is a new branch in the field of ML that can replace the manual feature extraction and selection with automatic learning of representative features [16]. Convolutional neural network (CNN) is one of the most effective DL models, which has also achieved a considerable performance in computer vision tasks and is also useful in transfer learning [17]. It is widely accepted that only the extracted features in the last convolutional layer are most suitable as the input vector of the classifier in most researches and applications of CNN models [18].

Since DL has the automatic feature extraction ability and ensemble learning can improve the accuracy and generalization performance of classifiers. This paper proposes a hybrid architecture that combines pre-trained DL techniques for feature extraction and random forest (RF) for binary classification of BC histological images over the BreakHis dataset at different magnification factors (MFs): 40X, 100X, 200X, and 400X. The main objective is to develop, evaluate and compare the hybrid RF models that are designed using: (1) three of the most recent DL techniques for feature extraction: DenseNet_201, MobileNet_V2, and Inception_V3 and (2) different number of trees (3, 5, 7 and 9).

Breiman [19] developed the concept of bagging in 1994 to improve classification by combining predictions of randomly generated training sets, the results proved that most of the improvement is obtained with unstable base models using only 10 bootstrap replicates, for that reason, the present study varies the number of trees between 3 and 9. The ML classifier that is frequently used to construct homogeneous ensembles is decision tree (DT) as reported in the systematic literature review [21], thus RF was used for classification. The three DL techniques (DenseNet_201, MobileNet_V2, and Inception_V3) were selected based on their performances as feature extractors (FE) for binary classification of breast pathological images over the BreakHis and FNAC

datasets [20]. All the empirical evaluations used: four classification performance criteria (accuracy, sensitivity, precision and F1-score), 5-fold cross-validation, Scott Knott (SK) statistical test to select the best cluster of the outperforming models, and Borda Count voting system to rank the best performing ones. The present study discusses four research questions (RQs):

- **(RQ1): Do hybrid RF models perform better than their trees?**
- **(RQ2): What is the best number of trees of the outperforming hybrid RF model for each FE?**
- **(RQ3): Do hybrid RF models perform better than the DT classifiers?**
- **(RQ4): What is the best hybrid RF model for each MF and over all the MFs?**

The main contributions of this paper are the following:

1. Designing three DT models using different DL techniques as FEs: DenseNet_201, Inception_V3, and MobileNet_V2.
2. Designing twelve hybrid RF models (4 models with a different number of trees for each DL architecture).
3. Evaluating the twelve hybrid RF models over the BreakHis dataset, considering the four MFs (40X, 100X, 200X and 400X) as four datasets.
4. Comparing the performances of the twelve hybrid RF models using SK clustering test and Borda Count voting method.
5. Comparing the performances of the twelve hybrid RF models with their base learners (trees) using SK clustering test and Borda Count voting method.
6. Comparing the performances of the best hybrid RF model for each classifier and FE with the DT classifier using SK clustering test and Borda Count voting method.

The remaining sections of this paper are provided as follows. Section 2 presents an overview of the three DL techniques used as FEs and the DT classifier; it also defines transfer learning, bagging in general and RF in particular. Section 3 describes the related work. Section 4 presents the materials and methods followed throughout this research. Section 5 presents and discusses the empirical results. Section 6 outlines the conclusion and future work.

2 Background

The present study uses three pre-trained DL techniques: DenseNet_201, Inception_V3 and MobileNet_V2 as FEs with RF as a classifier to design the hybrid architectures. They have been implemented using transfer learning for feature extraction and the homogeneous ensemble learning method: bagging for classification. This section gives a summary of the different methods and techniques used.

2.1 Transfer Learning

Transfer learning is a ML technique that transfer the knowledge of a model from a related source task to a second related task. Transfer learning only works in DL if the

model features learned from the first task are general [22]. This method optimizes its modeling by allowing rapid progress and improved performance. Compared with training from scratch, using a pre-trained convolutional neural network on a target dataset can significantly improve performance, while compensating for the lack of sufficient training data in the target task [23]. Transfer learning is widely used in computer vision, the models for image classification that result from a transfer learning approach based on pre-trained CNNs are usually composed of two parts: The convolutional base, which performs feature extraction, and the classifier, which classifies the input image based on the features extracted by the convolutional base. The typical way of conducting transfer learning with deep neural networks is usually expressed through the use of pre-trained models with 2 different methods:

- **Fine Tuning pre-trained models**: The idea of this method is to slightly adjust the parameters of a pre-trained model to allow knowledge transfer from the source task to the target task. Commonly used fine-tuning strategies include fine-tuning all the pre-trained parameters and fine-tuning few layers. The second strategy is made by unfreezing a few of the top layers of a frozen model base and jointly train both the newly added classifier layers and the last layers of the base model, in order to make it more relevant for the new specific task [24].
- **Pre-trained models as feature extractors**: This method uses the representations learned by a previous network to extract meaningful features from new samples. The main idea is to keep the convolutional base of the pre-trained model in its original form and then use its outputs to feed the classifier. This approach is much faster and give better results if the initial model was trained on a large benchmark dataset to solve a problem similar to the one in this research [25].

2.2 DL Techniques for Feature Extraction

Transfer learning with CNNs, well-trained on non-medical ImageNet dataset, has shown promising results for medical image analysis in recent years [26]. This study used the following pre-trained models for feature extraction:

- **DenseNet_201** is a convolutional neural network with 201 layers that connects each layer to every other layer in a feed-forward fashion. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters [27].
- **MobileNet_V2** is based on an inverted residual structure where the residual connections are between the bottleneck layers. The intermediate expansion layer uses lightweight depth wise convolutions to filter features as a source of non-linearity. As a whole, the architecture of MobileNetV2 contains the initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers [28].
- **Inception_V3** is the third version in a series of DL convolutional architectures. It is a CNN with 42 layers with a default input size fixed to 299 $\times$ 299 [29]. It has a block of parallel convolutional layer with a filter size of (3 $\times$ 3), followed by three inception modules with 288 filters each, 5 instances of the factorized inception modules with

768 filters, and two inception modules with 1280 filters. The outputs are concatenated and sent to a SoftMax output layer.

2.3 Decision Tree (DT)

DT is a flowchart-like tree structure that classifies instances by sorting them based on feature values, where each node in a DT represents a feature in an instance to be classified, and each branch represents a decision rule, and each leaf node represents the outcome. the topmost node is known as the root node and it is the feature that best divides the training data [30].

2.4 Bagging

Bagging or "Bootstrap aggregating" is a homogenous ensemble method that improves the accuracy by decreasing the variance and reducing the overfitting [31]. This method generates sample subsets (bootstrap) by randomly sampling from the training dataset with or without replacement, in the first case many of the original examples of the training dataset may be repeated while others may be left out. The base learners are then trained on the generated subsets and their predictions are combined to get an aggregated prediction. The aggregation averages over the base learners' results when predicting a numerical outcome and does a majority vote when predicting a class. The training is performed in a parallel manner and the accuracy increases if perturbing the learning set causes major changes in the predictor built [19, 32].

2.5 Random Forest

RF classifier is a bagging ensemble that produces multiple DTs that will be used to classify a new instance by the majority vote [33]. Each DT node uses a subset of attributes randomly selected from the whole original set of attributes, and each tree uses a different bootstrap sample data. RF classifier has been attracting increasing attention due to its excellent classification accuracy and high efficiency; It is a computationally efficient technique that can operate quickly over large datasets [34].

3 Related Work

Motivated by the success of DL in computer vision, various studies employed pre-trained DL models for binary classification of BC histopathological images using hybrid architectures (the feature and classification tasks are done with different models). For instance, Zerouaoui et Idri [20] developed and evaluated twenty-eight hybrid architectures combining seven recent DL 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 a binary classification of breast pathological images over the BreakHis and FNAC datasets. The hybrid architecture using the MLP classifier and DenseNet 201 for feature extraction was the top performing architecture

with higher accuracy values reaching 92.61%, 92%, 93.93%, and 91.73% over the four MF values of the BreakHis dataset: 40X, 100X, 200X, and 400X, respectively.

Kassani et al. [35] proposed ensemble model adapts three pre-trained CNNs: VGG19, MobileNet_V2, and DenseNet_201 to extract features and feed them into a multi-layer perceptron classifier. The proposed ensemble outperformed the single CNNs and the candidate ML classifiers, namely, DT, RF, XGBoost, AdaBoost and bagging.

Saxena et al. [36] investigated ten different pre-trained CNNs (AlexNet_43, VGG_16 and VGG_19; GoogLeNet, ResNet18, ResNet_50, ResNet_101, Inception_V3, Inception_ResNet_V2 and SqueezeNet) for extracting the features from BC histopathology images from the BreakHis dataset. For classification, they examined different classifiers which are: SVM, KNN, DT, RF, and boosting. They selected SVM with linear kernel for the classification since it performed better than the other classifiers. However, the selection of the classifier was only performed with ResNet_50 on MF 40X. The best performing models used SVM as classifier and as FE: ResNet50 for MF 40X, ResNet101 for MF 100X, ResNet50 for MF 200X, and AlexNet for MF 400X with an accuracy of 87.71%, 88.92%, 90.12% and 85.16% respectively.

4 Materials and Methods

This section presents the data preparation process, the experimental process followed to carry out all the empirical evaluations, the threats of validity and the abbreviations.

4.1 Data Preparation

The data preparation process followed for the BreakHis dataset in the current study is similar to the one that has been followed in [37]; thereafter we summarize the different steps of this process. The present study used intensity normalization to ensure that each input parameter (pixel, in this case) has a similar data distribution and to remove highlighted regions and shadows [38]. The input images were normalized to the standard normal distribution using min-max. Furthermore, before feeding input images into the proposed models, Contrast Limited Adaptive Histogram Equalization (CLAHE) is a necessary step to improve the contrast in images [39]. For the BreakHis dataset, 70% of the histological images represent the malignant class. To overcome this limitation, data augmentation [40] was used to resample the BreakHis dataset after the pre-processing step to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. After using data augmentation for the original BreakHis dataset, a new balanced dataset was obtained where the total number of benign samples was increased by 2 times to obtain the same number of malignant images.

4.2 Experimental Process

In this experiment, the models are evaluated using k-folds cross validation with $k = 5$.

The empirical evaluations use RF, three DL techniques as FE: DenseNet_201, Inception_V3 and MobileNet_V2, and the BreakHis dataset which contains images with various MFs including 40X, 100X, 200X, and 400X. Moreover, the SK test [41] based

on accuracy was used to cluster the models in order to identify the best SK clusters (i.e. containing the best models that are statistically indifferent in terms of accuracy). Thereafter, Borda Count [42] was used to rank the models belonging to each best SK cluster using accuracy, recall, precision and F1-score. Figure 1 shows the methodology followed to carry out all the empirical evaluations.

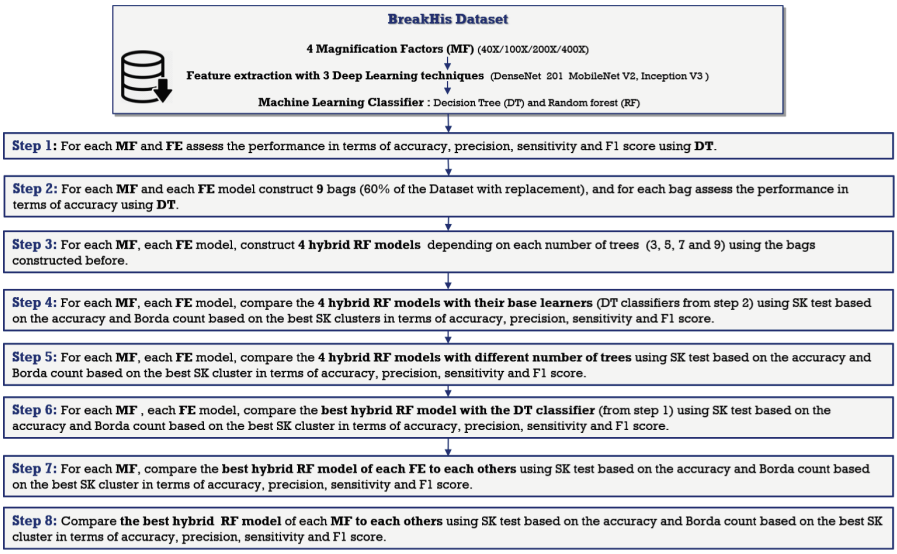


Fig. 1. Experimental process.

4.3 Threats of Validity

This paper used the k-fold cross validation as an internal threat with $k = 5$, The choice of k is usually 5 or 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance [43]. Transfer learning was also used to extract the features of the images with the most powerful DL techniques. For the external validity, this paper used the BreakHis dataset which contains histological images with four different MFs. Each MF can be considered as a dataset, which was helpful to compare the performance of the different models and FE for each MF. However, it would be interesting to study if the robustness of the models holds with different BC image datasets. For the reliability of the classifier performances obtained, this paper focused on four classification metrics (accuracy, recall, precision and F1-Score) to construct validity. The main reasons behind the choice of these performance criteria are: (1) most of the studies used them to measure the classification performance, 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.

4.4 Abbreviation

In order to assist the reader and shorten the names of the models, this paper uses the following naming rules. The abbreviation of the hybrid RF models is RF, plus the abbreviation of the feature extractor, followed by the number of trees. The name of the FE technique used with the RF model is abbreviated using the first letter of the DL model (D for DenseNet_201, I for Inception_V3 and M for MobileNet_V2). As an example: RFD9 is the hybrid RF model of 9 trees constructed with DenseNet_201 as FE.

In the next section, the DL techniques were abbreviated using their first 3 letters (DEN for DenseNet_201, INC for Inception_V3 and MOB for MobileNet_V2) for tables and figures.

5 Results and Discussion

This section presents and discusses the results of the empirical evaluations. Python 3 was used with the two DL frameworks Keras and Tensorflow as DL backend and with OpenCV for image pre-processing. The training and testing steps are running using Google Colaboratory.

5.1 Evaluating and Comparing the Hybrid RF Models with Their Trees

This subsection evaluates and compares the hybrid RF models with their base learners (trees). For each MF and FE, the hybrid RF models were first compared in terms of accuracy with the trees that construct them. Thereafter, the SK statistical test was used to determine the best cluster, and finally Borda Count was used to rank the models belonging to the best SK cluster. Figures 2 shows the difference in % between the accuracy value of hybrid RF models and the mean accuracy value of their trees over the 4 different MFs. The number 3, 5, 7 or 9 refers to the number of trees that construct each RF model.

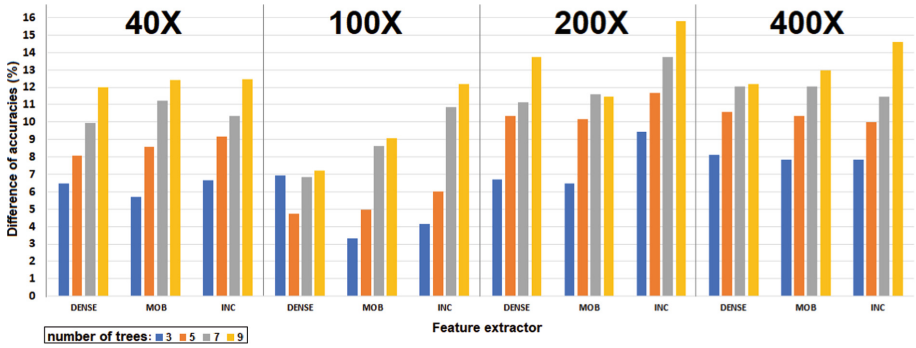


Fig. 2. Difference between the accuracy of the hybrid RF models and their trees over each MF and FE.

For each MF and FE, the observations of the results are as follow:

- The accuracy of the hybrid RF models is always better than the mean accuracy of their trees.
- The difference of accuracies increases with the number of trees that construct them (15.80% for the highest difference value with a RF of 9 trees, Inception_V3 and MF 200X and 3.31% for the lowest one with a RF of 3 trees, MobileNet_V2 and 100X).

Thereafter, the SK statistical test was used to determine the best cluster, it was found that for the majority of MFs and FEs two clusters were obtained; the best cluster contains the hybrid RF models while the other cluster contains all the trees. However, there are two exceptions for MF 100X and the hybrid RF model constructed using three trees:

- When using DenseNet_201 as FE, two clusters were obtained, the best one contains the hybrid RF model and one of its trees, while the other cluster contains the two other trees. Which means that statistically the hybrid RF model and one of its trees show the same predictive capabilities in terms of accuracy.
- When using MobileNet_V2 as FE, only one cluster was obtained, which means that statistically the hybrid RF model and its trees show the same predictive capabilities.

In order to identify which model is the best for MobileNet_V2 and DenseNet_201 as FE with MF 100X, the models of the best SK cluster of each MF and FE were ranked by using the Borda Count voting system based on accuracy, sensitivity, F1-score and precision. The results of the Borda Count voting method show that the hybrid RF models are always ranked first.

To conclude, hybrid RF models outperform their trees and the difference of accuracies between them and their trees is relatively important.

5.2 Evaluating and Comparing the Impact of the Number of Trees on the Hybrid RF Models Performances Over Each FE and MF

This subsection evaluates and compares the hybrid RF models implemented with the same FE but with a different number of trees (3, 5, 7 or 9) for each MF, in order to identify the number of trees that gives the best performance. The performances of the hybrid RF models were evaluated using four criteria: accuracy, precision, recall and F1-score. For each FE and each MF, the hybrid RF models were first compared in terms of accuracy.

Figure 3 shows the comparison of the mean accuracy values of the hybrid RF model using the four number of trees, the 3 DL techniques for FE over each MF value: 40X, 100X, 200X and 400X. It can be seen that:

- The best accuracy values were obtained when using the DenseNet_201 as FE for the MF 100X, and with Inception_V3 for the MF 40X, 200X and 400X. The worst accuracy values were obtained when using MobileNet_V2 for MF 40X, 200X and 400X, and when using Inception_V3 for MF 100X.
- The best accuracy values when using the DenseNet_201 as FE with the MF 100X reached 82.57% with a hybrid RF model of 9 trees, and with Inception V2 as FE the values were 84.31%, 85.88% and 82.35% with a hybrid RF model of 9 trees for the MF 100X, 200X and 400X respectively.

- The worst accuracy values when using the MobileNet_V2 were: 76.79%, 77.49% and 72.15% with a hybrid RF model of 3 trees for the MF 40X, 200X and 400X respectively. For MF 100X, the worst accuracy values when using the Inception_V3 was: 76.60%.

Thereafter, the SK statistical test was used to determine the best cluster, and finally Borda Count was used to rank the hybrid RF models belonging to the best SK cluster. It was found that the ranks are ascending: the hybrid RF models with more trees perform better than the others; but there are two exceptions. The hybrid RF model of 9 trees is outperformed by the hybrid RF model of 7 trees when using DenseNet_201 and MobileNet_V2 for MF 200X and MF 400X respectively.

In summary, (1) the hybrid RF models of 9 trees were generally ranked first, and (2) the hybrid RF models of 3 trees are always ranked last.

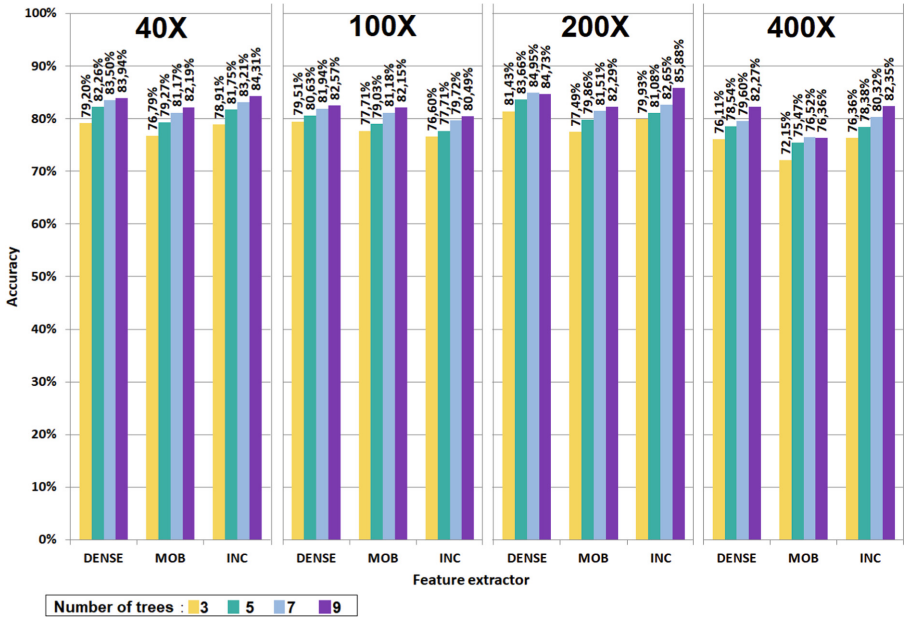


Fig. 3. Mean accuracy values of the hybrid RF models using different number of trees over each MF and FE.

5.3 Evaluating and Comparing the Best Hybrid RF Models and DT Classifiers

This subsection evaluates and compares the first ranked hybrid RF model found in RQ2 with the DT classifier over each MF and FE. Table 1 summarizes the testing accuracy values of the best hybrid RF model for each FE and the DT classifier designed using the same FE. Two different clusters were obtained for the results of SK test for the hybrid RF model and DT classifier over each FE and MF, where the first cluster was always represented by the hybrid RF model and the second one with the DT classifier. Therefore, the hybrid RF model outperforms the DT classifier for each MF and FE.

Table 1. Mean accuracy values of the best hybrid RF model and the DT classifier over each MF and FE.

MF \ FE	Hybrid RF accuracy (%)			DT accuracy (%)		
	DEN	**MOB**	**INC**	**DEN**	**MOB**	**INC**
40X	83.94	82.19	**84.31**	79.34	75.26	74.31
100X	**82.57**	80.49	82.15	76.39	76.39	74.31
200X	84.95	82.29	**85.88**	74.12	74.41	74.12
400X	82.27	76.52	**82.35**	73.20	70.04	72.55

5.4 Comparing the Best Hybrid RF Models of Each FE for Each MF and Over all the MFs

This subsection evaluates and compares the best hybrid RF models of the three FE for each MF and over all the MFs. Figure 4 shows the results of the SK test of the best hybrid RF models for each FE over each MF. It is observable that one cluster was obtained for MF 40X, 100X and 200X and two clusters for MF 400X.

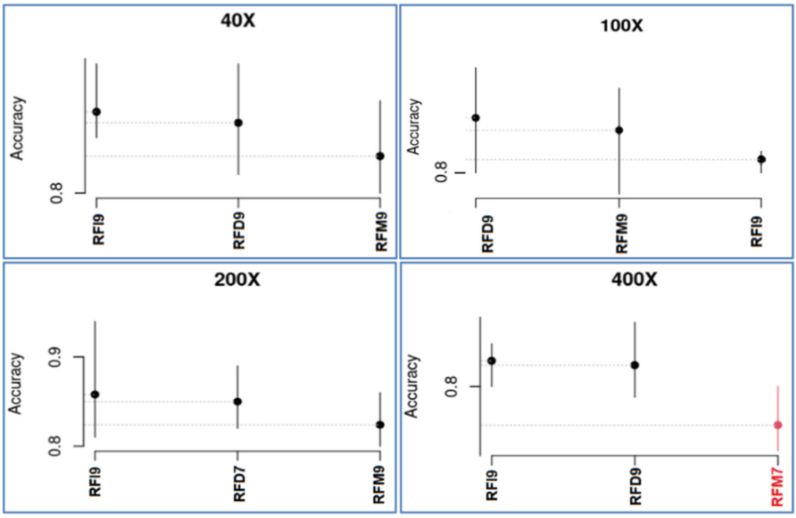


Fig. 4. SK Results of the best hybrid RF models for each FE over each MF

The hybrid RF models of the best SK cluster of each MF were ranked by using the Borda Count voting system based on accuracy, sensitivity, F1-score and precision. After using the Borda Count method it was found that:

- For MF 40X, the best hybrid RF model is constructed using Inception_V3 as FE with 9 trees. The second ranked hybrid RF model is constructed using DenseNet_201 as FE with 9 trees. The third one has 9 trees and was constructed with MobileNet_V2 as FE.

- For MF 100X, the best hybrid RF model is constructed using DenseNet_201 as FE with 9 trees. The second ranked hybrid RF model is constructed using MobileNet_V2 as FE with 9 trees. The third one has 9 trees and was constructed with Inception_V3 as FE.
- For MF 200X, the best hybrid RF model is constructed using Inception_V3 as FE with 9 trees. The second ranked hybrid RF model is constructed using DenseNet_201 as FE with 7 trees. The third one has 9 trees and was constructed with MobileNet_V2 as FE.
- For MF 400X, the best hybrid RF model is constructed using Inception_V3 as FE with 9 trees. The second ranked RF hybrid RF model is constructed using DenseNet_201 as FE with 9 trees. The third one has 7 trees and was constructed with MobileNet_V2 as FE.

Table 2 summarizes the three best hybrid RF models for each MF. It was found that the best ranked hybrid RF models always have 9 trees. For the feature extraction techniques: Inception_V3 was ranked first the most, DenseNet_201 was ranked second the most and MobileNet_V2 was ranked third the most.

Table 2. Best hybrid RF models for each MF

MF	1st ranked	2nd ranked	3rd ranked
40X	RFI9	RFD9	RFM9
100X	RFD9	RFM9	RFI9
200X	RFI9	RFD7	RFM9
400X	RFI9	RFD9	RFM7

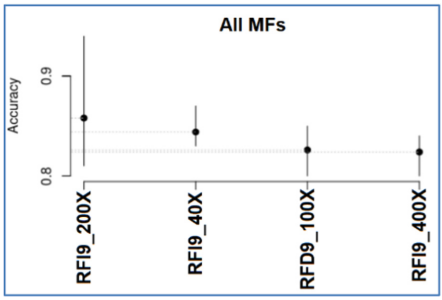


Fig. 5. SK Results of the best hybrid RF models of each MF

Figure 5 shows the results of SK test for the best hybrid RF models over the four MFs. Only one cluster was obtained that contains all the models, which means that the best hybrid RF models for each MF are statistically indifferent in terms of accuracy. In order to identify which hybrid RF model is the best over all the MFs, the hybrid RF models of the best SK cluster were ranked by using the Borda Count voting system based

on accuracy, sensitivity, F1-score and precision. Table 3 presents the ranking results of the best hybrid RF models using the Borda count voting method based on accuracy, precision, recall, and F1-score. It was found that the best hybrid RF model over all the MFs uses MF 200X, the second-best model uses MF 40X, and the third-best model uses MF 100X.

Table 3. Best hybrid RF models over all the MFs

Model	MF	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)	Rank
RFI9	200X	85.88	85.48	87.03	86.24	**1**
RFI9	40X	84.31	84.07	86.35	85.13	**2**
RFD9	100X	82.57	81.54	83.47	82.47	**3**
RFI9	400X	82.35	82.07	83.62	82.78	**4**

To sum up, the best hybrid architecture regardless of the MF is designed using the hybrid RF model with 9 trees as a classifier and Inception_V3 as FE, the second-best one is constructed using the hybrid RF model with 9 trees as a classifier and Inception_V3 as FE, the third best one is constructed using the hybrid RF model with 9 trees as a classifier and DenseNet_201 as FE. The fourth ranked one is designed using the hybrid RF model with 9 trees as classifier and Inception_V3 as FE.

6 Conclusion and Future Work

This paper presented and discussed the results of an empirical comparative study of forty-eight ($48 = 4$ MF $\times$ 3 DL techniques $\times$ 4 number of trees) hybrid RF models for BC histopathological image classification. This study used hybrid RF models for classification with different number of trees (3, 5, 7 and 9) and three DL techniques as FEs (DenseNet_201, Inception_V3 and MobileNet_V2). 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 models over the BreakHis dataset with its four MF (40X, 100X, 200X and 400X). The results showed that the hybrid RF models with the highest number of trees consistently perform well over all the MFs in terms of their accuracy, precision, recall and F1-score, and they also outperform the trees that construct them and the DT classifiers. As a result, this paper suggests that RF could be considered for the task of BC histological imaging classification with Inception_V3 as FE. Ongoing work focuses on implementing hybrid bagging ensembles using other classifiers as base learners to possibly improve the accuracy and comparing them with the hybrid RF models.

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.

References

1. Breast Cancer Facts and Statistics. https://www.breastcancer.org/facts-statistics. Accessed 08 Apr 2022
2. Ginsburg, O., et al.: Breast cancer early detection: A phased approach to implementation. Cancer **126**, 2379–2393 (2020). https://doi.org/10.1002/cncr.32887
3. Yassin, N.I.R., Omran, S., El Houby, E.M.F., Allam, H.: Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. Comput. Methods Programs Biomed. **156**, 25–45 (2018). https://doi.org/10.1016/j.cmpb.2017.12.012
4. Abdar, M., et al.: A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recogn. Lett. **132**, 123–131 (2020). https://doi.org/10.1016/j.patrec.2018.11.004
5. Hamed, G., Marey, M.A.E.-R., Amin, S.E.-S., Tolba, M.F.: Deep learning in breast cancer detection and classification. In: Hassanien, A.-E., Azar, A.T., Gaber, T., Oliva, D., Tolba, F.M. (eds.) AICV 2020. AISC, vol. 1153, pp. 322–333. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-44289-7_30
6. Ho, T.K.: Multiple classifier combination: lessons and next steps. In: Bunke, H., Kandel, A. (eds.) Hybrid Methods in Pattern Recognition, pp. 171–198. WORLD SCIENTIFIC (2002). https://doi.org/10.1142/9789812778147_0007
7. Kuncheva, L.I.: Combining Pattern Classifiers, p. 382 (2014)
8. Polikar, R.: Ensemble learning. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning, pp. 1–34. Springer, Boston (2012). https://doi.org/10.1007/978-1-4419-9326-7_1
9. Sagi, O., Rokach, L.: Ensemble learning: a survey. WIREs Data Min. Knowl. Discov. **8**, e1249 (2018). https://doi.org/10.1002/widm.1249
10. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. JAIR **11**, 169–198 (1999). https://doi.org/10.1613/jair.614
11. Oza, N.C., Tumer, K.: Classifier ensembles: select real-world applications. Inf. Fus. **9**, 4–20 (2008). https://doi.org/10.1016/j.inffus.2007.07.002
12. Kuncheva, L.I.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy, p. 27 (2003)
13. Brown, G., Kuncheva, L.I.: "Good" and "Bad" diversity in majority vote ensembles. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 124–133. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12127-2_13
14. El Ouassif, B., Idri, A., Hosni, M.: Investigating accuracy and diversity in heterogeneous ensembles for breast cancer classification. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12950, pp. 263–281. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86960-1_19
15. Wang, G., Hao, J., Ma, J., Jiang, H.: A comparative assessment of ensemble learning for credit scoring. Expert Syst. Appl. **38**, 223–230 (2011). https://doi.org/10.1016/j.eswa.2010.06.048
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015). https://doi.org/10.1038/nature14539
17. del Rio, F., Messina, P., Dominguez, V., Parra, D.: Do Better ImageNet Models Transfer Better... for Image Recommendation? arXiv:1807.09870 [cs] (2018)
18. Xu, G., Liu, M., Jiang, Z., Söffker, D., Shen, W.: Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. Sensors **19**, 1088 (2019). https://doi.org/10.3390/s19051088
19. Breiman, L.: Bagging predictors. Mach Learn. **24**, 123–140 (1996). https://doi.org/10.1007/BF00058655
20. Zerouaoui, H., Idri, A.: Deep hybrid architectures for binary classification of medical breast cancer images. Biomed. Signal Process. Control **71**, 103226 (2022). https://doi.org/10.1016/j.bspc.2021.103226

21. Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J.M., Fernández Alemán, J.L.: Reviewing ensemble classification methods in breast cancer. Comput. Methods Prog. Biomed. **177**, 89–112 (2019). https://doi.org/10.1016/j.cmpb.2019.05.019
22. Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., Feris, R.: SpotTune: Transfer Learning Through Adaptive Fine-Tuning, p. 10 (2018)
23. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**, 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
24. Alshalali, T., Josyula, D.: Fine-tuning of pre-trained deep learning models with extreme learning machine. In: 2018 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 469–473. IEEE, Las Vegas, NV, USA (2018). https://doi.org/10.1109/CSCI46756.2018.00096
25. Ahmed, A., Yu, K., Xu, W., Gong, Y., Xing, E.: Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 69–82. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_6
26. Morid, M.A., Borjali, A., Del Fiol, G.: A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. **128**, 104115 (2021). https://doi.org/10.1016/j.compbiomed.2020.104115
27. Wang, S.-H., Zhang, Y.-D.: DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. ACM Trans. Multimedia Comput. Commun. Appl. **16**, 1–19 (2020). https://doi.org/10.1145/3341095
28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520. IEEE, Salt Lake City, UT (2018). https://doi.org/10.1109/CVPR.2018.00474
29. Inception V3 Deep Convolutional Architecture For Classifying Acute. https://www.intel.com/content/www/us/en/develop/articles/inception-v3-deep-convolutional-architecture-for-classifying-acute-myeloidlymphoblastic.html. Accessed 3 June 2021
30. Iqbal, M., Yan, Z.: Supervised machine learning approaches: a survey. Int. J. Soft Comput. **5**, 946–952 (2015). https://doi.org/10.21917/ijsc.2015.0133
31. Liang, G., Zhu, X., Zhang, C.: An Empirical Study of Bagging Predictors for Different Learning Algorithms, p. 2 (2011)
32. Bühlmann, P., Yu, B.: Analyzing bagging. Ann. Statist. **30** (2002). https://doi.org/10.1214/aos/1031689014
33. Adele Cutler, D., Cutler, R., Stevens, J.R.: Random forests. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning, pp. 157–175. Springer, Boston (2012). https://doi.org/10.1007/978-1-4419-9326-7_5
34. Oshiro, T.M., Perez, P.S., Baranauskas, J.A.: How many trees in a random forest? In: Perner, P. (ed.) MLDM 2012. LNCS (LNAI), vol. 7376, pp. 154–168. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31537-4_13
35. Kassani, S.H., Kassani, P.H., Wesolowski, M.J., Schneider, K.A., Deters, R.: Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks. arXiv:1909.11870 [cs, eess] (2019)
36. Saxena, S., Shukla, S., Gyanchandani, M.: Pre-trained convolutional neural networks as feature extractors for diagnosis of breast cancer using histopathology. Int. J. Imaging Syst. Technol. **30**, 577–591 (2020). https://doi.org/10.1002/ima.22399
37. Zerouaoui, H., Idri, A., Nakach, F.Z., Hadri, R.E.: Breast fine needle cytological classification using deep hybrid architectures. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12950, pp. 186–202. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86960-1_14
38. Nikhil, B.: Image Data Pre-Processing for Neural Networks. https://becominghuman.ai/image-data-pre-processing-for-neural-networks-498289068258. Accessed 12 May 2021

39. Yussof, W.: Performing Contrast Limited Adaptive Histogram Equalization Technique on Combined Color Models for Underwater Image Enhancement (2013)
40. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019). https://doi.org/10.1186/s40537-019-0197-0
41. ScottKnott: a package for performing the Scott-Knott clustering algorithm in R. https://www.scielo.br/scielo.php?script=sci_arttext&pid=S2179-84512014000100002. Accessed 20 May 2021
42. Borda Count | Mathematics for the Liberal Arts. https://courses.lumenlearning.com/waymakermath4libarts/chapter/borda-count/. Accessed 21 May 2021
43. Hastie, T., Tibshirani, R., Friedman, J.: Ensemble learning. In: Hastie, T., Tibshirani, R., Friedman, J. (eds.) The Elements of Statistical Learning. SSS, pp. 605–624. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7_16