International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 4, August 2024, pp. 4261~4272
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i4.pp4261-4272

Optimizing breast cancer diagnosis: combining hybrid architectures through Apache Spark

Chaymae Taib[1], Otman Abdoun[2], El Khatir Haimoudi[1]
[1]Computer Science Department, Polydisciplinary Faculty, Abdelmalek Essaadi University, Larache, Morocco
[2]Computer Science Department, Faculty of Science, Abdelmalek Essaadi University, Tetouan, Morocco

Article Info

Article history:

Received Dec 25, 2023
Revised Mar 27, 2024
Accepted Apr 2, 2024

Keywords:

Big data
Deep learning
Hybrid architectures
Machine learning
Pre-processing
Spark

ABSTRACT

Early detection and diagnosis of breast cancer are critical for saving lives. This paper addresses two major challenges associated with this task: the vast amount of data processing involved and the need for early detection of breast cancer. To tackle these issues, we developed thirty hybrid architectures by combining five deep learning techniques (Xception, Inception-V3, ResNet50, VGG16, VGG19) as feature extractors and six classifiers (random forest, logistic regression, naive Bayes, gradient-boosted tree, decision tree, and support vector machine) implemented on the Spark framework. We evaluated the performance of these architectures using four classification criteria. The results, analyzed using Scott Knott's statistical test, demonstrated the effectiveness of merging deep learning feature extraction techniques with traditional classifiers for classifying breast cancer into malignant and benign tumors. Notably, the hybrid architecture using logistic regression as the classifier and ResNet50 for feature extraction (RESLR) emerged as the top performer. It achieved impressive accuracy scores of 98.20%, 96.59%, 96.64%, and 94.84% across the Break-His dataset at different magnifications (40X, 100X, 200X, and 400X) respectively. Additionally, RESLR achieved an accuracy of 97.05% on the ICIAR dataset and a remarkable accuracy of 95.31% on the FNAC dataset.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Chaymae Taib
Computer Science Department, Polydisciplinary Faculty, Abdelmalek Essaadi University
Larache, Morocco
Email: taib.chaymae@etu.uae.ac.ma

1. INTRODUCTION

The sixth-leading cause of cancer-related deaths globally, and the primary cause among women, is breast cancer, projected to be the most frequently diagnosed cancer worldwide by 2020. Approximately 2.26 million new cases of breast cancer are anticipated, with an estimated 685,000 deaths [1]. Recognized as a significant global health issue, cancer poses substantial challenges. The global burden of disease (GBD) estimates that in 2020 alone, there will be approximately 19.3 million new cancer cases and nearly 10 million cancer-related deaths. Among various cancer types, breast cancer stands out as the most common, significantly contributing to female mortality globally. In 2020, an estimated 2.3 million women were affected by this disease, highlighting its widespread impact.

As the number of women afflicted with breast cancer increased, radiologists faced challenges in handling timely diagnoses [2]. While the touchstone standard for breast cancer diagnosis remains pathological examination, it is time-consuming and resource-intensive [3]. Due to the superficial nature of breasts, imaging technologies can detect anomalies in breast mass, proving to be valuable. Image analysis in medicine, particularly using deep learning, has shown superior performance in various domains [4], including classification [5], [6], detection, and segmentation [7].

Deep learning techniques offer exceptional outcomes [8], [9] aiding radiologists in making informed decisions and enabling early detection [9]. These algorithms automatically extract features, retrieve information from data, and learn advanced abstract data representations [10]. However, classical machine learning techniques, while producing accurate results for breast cancer classification [11], require less time in training and parameter tuning. Numerous studies have been conducted to evaluate tumor detections by using traditional deep learning techniques, using the IRMA dataset, visual geometry group 16-layer model (VGG16) and residual network with 50 layers (ResNet50) were applied to classify normal and abnormal tumors. Results indicate VGG16 achieved superior accuracy at 94%, surpassing ResNet50's 91.7% [12]. Results indicate that K-nearest neighbors (KNN) achieves the highest accuracy (97.51%) and the lowest error rate, outperforming the naive Bayes (NB) classifier (96.19%) [13]. Therefore, the researchers have developed hybrid architectures effectively combining deep learning techniques for feature extraction with traditional machine learning methods for classification [14]–[18].

The multi-layer perceptron and DenseNet 201 (MDEN) architecture, combining the multi-layer perceptron (MLP) classifier and DenseNet 201 for feature extraction, emerged as the top-performing model with up to 99% accuracy on the Fine needle aspiration cytology (FNAC) dataset [19]. Its success lies in the synergistic combination of dense connections from DenseNet and the multilayer perceptron's classification capabilities, which effectively captured intricate patterns within the data. On the other hand, the convolutional neural network-long short-term memory (CNN-LSTM) model achieved a remarkable accuracy of 96% by combining convolutional neural network (CNN) for feature extraction and long short-term memory (LSTM) for classification, showcasing the effectiveness of leveraging both convolutional and recurrent neural network architectures for complex data analysis tasks [20].

In addressing the categorization challenge within the Wisconsin diagnostic breast cancer (WDBC) dataset, researchers employed an ensemble of support vector machines (SVMs) that demonstrated exceptional performance. This ensemble achieved over 99% accuracy in identifying test data, and notably, 100% accuracy in predicting benign tumors [21]. Moreover, in another study focused on lung cancer classification, a hybrid approach utilizing SVM and neural networks was adopted. This approach yielded even higher results, with an average precision value of 98.17% [22].

The rising incidence of diseases such as breast cancer has resulted in a surge in the volume of medical images, presenting challenges for conventional deep learning and machine learning techniques. In response, big data has emerged as a promising solution. Studies have shown that leveraging big data analytics has led to enhancements in decision-making quality [23], [24]. This highlights the potential of big data in addressing the complexities posed by the growing number of medical images.

In the realm of classifying extensive datasets, researchers have developed a distributed heterogeneous boosting-inspired ensemble classifier (DHBoost). This classifier, which operates on the MapReduce computing paradigm and Apache Spark framework, has demonstrated superior performance compared to Spark ensemble classifiers (such as Spark-random forests (Spark-RF) and Spark-gradient-boosted trees (Spark-GBT)) across various scenarios [25]. The efficacy of DHBoost underscores the importance of leveraging innovative approaches, particularly within the context of large-scale data processing.

In this study, our primary objective is to identify an optimal combined architecture that attains superior accuracy within a concise timeframe. To achieve this goal, we thoroughly investigate thirty hybrid architectures, utilizing a diverse set of six classifiers and incorporating five deep learning techniques. Our experimentation is conducted across three distinct datasets for binary breast cancer classification. The implementation is carried out on the Apache Spark framework, strategically chosen to enhance both execution speed and training efficiency. This comprehensive approach is designed to tackle two critical challenges: effectively managing substantial volumes of data and facilitating early detection. The primary objective of this study is to examine and provide insights into five fundamental aspects (FAs):

- (FA1): To what extent do the thirty hybrid architectures developed on Apache Spark demonstrate performance in classifying breast cancer?
- (FA2): Is there a feature extractor approach that obviously exceeds others when used in a heterogeneous design?
- (FA3): Is there a hybrid design that obviously outperforms others, irrespective of the selection of feature extractor and classifier?
- (FA4): Is there a combination design that clearly outperforms others, independent of the dataset?
- (FA5): Did the implementation of Apache Spark have a noticeable impact on the outcomes, indicating an improvement?

The structure of this work unfolds as follows: section 2, labeled research method, delves into the dataset, preprocessing methods, evaluation metrics, and abbreviations. In section 3, the proposed method employed in this study is detailed. Empirical findings are succinctly summarized and analyzed in section 4. Moving forward, section 5 engages in a thorough discussion of the results while also outlining potential avenues for future research projects.

2. RESEARCH METHOD

2.1. The dataset description and preprocessing

Table 1 offers a detailed overview of the datasets utilized in the study, including references for the ICIAR, FNAC, and BreakHis datasets, along with their respective types and sizes. Specifically, the ICIAR dataset comprises 400 microscope images, while the FNAC dataset includes 113 malignant and 99 benign cases. Moreover, the BreakHis dataset encompasses 7,909 images distributed across four magnification factors: 40X, 100X, 200X, and 400X.

Table 1. Description and preprocessing of three datasets

DATASET	Type	Size	Preprocessing of data
ICIAR [26]	Images	400 microscope images	Rotation: Apply a random rotation to the image within a range of 30 degrees. Zoom: Adjust the image by zooming out with a range value of 0.2. Apply shearing: Distort the image by shifting one part in one direction and the other part in the opposite direction, within a range of 0.2. Adjust width and height: Shift the image both horizontally and vertically, with a range set at 0.3. Implement horizontal flipping: Create a mirror image along the horizontal axis. Utilize fill mode: Substitute empty areas with the nearest pixel value. Rescale: Adjust pixel values to a new range of 0-1 from the original 0-255 range.
FNAC [27]	Images	113 malignant and 99 benign cases	
BreakHis [28]	Images	7,909 images across four magnification factors 40X, 100X, 200X, and 400X	

2.2. Evaluation metrics

In evaluating the efficacy of the heterogeneous methods, we employed a set of performance metrics. These metrics, which encompass accuracy, precision, recall, and F1-score, were utilized to assess the models' performance. The equations for calculating these metrics are presented as (1)-(4).

$$Accuracy = (TP + TN)/(TN + TP + FP + FN)) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN)) \tag{3}$$

$$F1 = 2 \times (Recall \times precision)/(Recall + precision) \tag{4}$$

2.3. Scott Knott statistical tests

The Scott-Knott (SK) technique, introduced by Scott and Knott in 1974, serves as a valuable exploratory clustering method in the analysis of variance (ANOVA) domain. The primary objective of the SK method is to discern overlapping groups by conducting multiple comparisons of treatment means. Renowned for its simplicity and robustness, this method is commonly employed as a hierarchical clustering algorithm [29]. In this study, the Scott-Knott statistical test was applied to the accuracy results acquired [30].

2.4. Abbreviation

In Table 2, the naming conventions applied to abbreviate the hybrid architectures are presented. These guidelines serve to provide a standardized and concise representation of the various hybrid models utilized in the study. By adhering to these conventions, researchers can easily identify and reference the specific architectures employed.

Table 2. Hybrid architecture abbreviation guidelines table

Abbreviation	Description	Abbreviation	Description
LR	Logistic regression	NBXCP	NB with Xception
SVM	Support vector machine learning	NBINCP	NB with Inception-V3
RF	Random forest	RFXCP	RF with Xception
NB	Naïve Bayes	NBV16	NB with VGG16
GBT	Greed boosting tree	RESRF	RF with ResNet50
RESLR	logistic regression with ResNet50	RFINCP	RF with Inception-V3
SVMINCP	SVM with Inception-V3	RFV16	RF with VGG16
SVMRES	SVM with ResNet50	GBTRES	GBT with ResNet50
LRINCP	logistic regression with Inception-V3	GBTXCP	GBT with Xception
SVMV16	SVM with VGG16	DTRES	DT with ResNet50
LRV19	Logistic regression with VGG19	GBTV19	GBT with VGG19
SVMXCP	SVM with Xception	DTV19	DT with VGG19
SVMV19	SVM with VGG19	RFV19	RF with VGG19
LRXCP	logistic regression with Xception	DTXCP	DT with Xception
NBV19	NB with VGG19	DTINCP	DT with Inception-V3
DTV16	Decision tree with VGG16	GBTV16	GBT with VGG16
LRV16	logistic regression with VGG16	GBTINCP	GBT with Inception-V3
NBRES	NB with ResNet50	-	-

3. PROPOSED METHOD

The proposed methodology involves the creation of hybrid architectures, employing six classifiers and five pre-trained CNN architectures for feature extraction in breast cancer histopathological images using the spark framework. Harnessing the capabilities of Apache Spark, particularly its spark deep learning pipeline, the approach incorporates utility functions to efficiently manage large-scale datasets. It seamlessly loads millions of images into a distributed spark dataframe, utilizing automatic decoding techniques for parallel and efficient processing across the spark cluster, enabling extensive manipulation of the data. The workflow, illustrated in Figure 1, delineates the steps followed in this experiment.

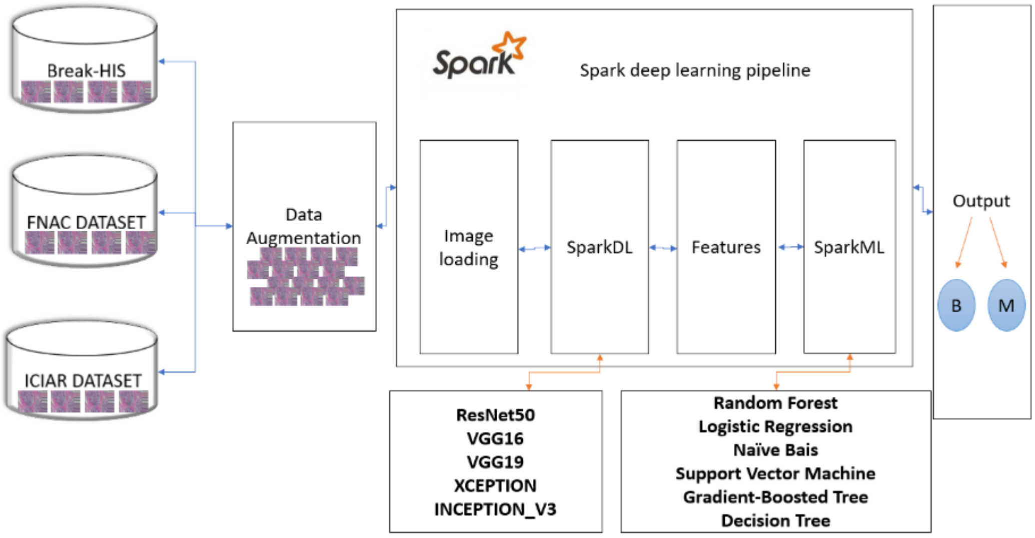

Figure 1. The overview of our process

In this proposal, 30 hybrid architectures (HA) were developed for each dataset (magnification factors (MF) at 40X, 100X, 200X, and 400X magnifications). All architectures provided in SparkDL from the DataBricks library (VGG16, Inception-V3, Xception, VGG19, ResNet50) were utilized for feature extraction. The extracted features were then fed into six classifiers (GBT, SVM, RF, LR, NB, DT). Specific configurations were employed for building and training the proposed models:

- The input images for the BreakHis dataset were standardized to a 512×512 pixel size after preprocessing.
- Transfer learning techniques were applied to all feature extraction methods (VGG16, ResNet50, Inception-V3, VGG19, Xception). This involved freezing the convolutional base of each architecture, and ImageNet weights were downloaded and used as kernel weights for extracting features.

- All the six machine learning classifiers (GBT, SVM, RF, LR, NB, and DT) used to classify the histopathological images into malignant or begin. These classifiers were used without tuning using the default parameters.

After training all thirty models, we conduct a thorough evaluation of their performance, taking into account metrics such as precision, accuracy, F1-score, and recall. To assess the significance of accuracy values among these models, we utilize the Scott-Knott (SK) statistical test for comparison. This comprehensive analysis allows for a nuanced understanding of the efficacy and relative performance of each model.

4. RESULTS AND DISCUSSION

This section presents the empirical findings derived from the evaluation of hybrid architectures across three diverse datasets (Break-His, ICIAR, FNAC). The assessment places emphasis on four fundamental performance measures. Initially, it assesses each classifier’s accuracy (FA1). Subsequently, it investigates the influence of five deep learning feature extraction methods on the four classifiers to determine those that contribute positively to classification performance (FA2). The subsequent analysis involves a comparison of the best-performing HB among the six classifiers to determine the optimal design (FA3). Additionally, a comprehensive evaluation of the thirty developed architectures, regardless of the dataset used, is conducted (FA4). Lastly, the study examines the performance improvement achieved by incorporating Apache Spark into the experimental workflow (FA5).

All empirical assessments were conducted in Python, utilizing the Keras and TensorFlow deep learning frameworks. The experiments were executed on a graphics processing unit (GPU) processing unit with 8 cores, 25 GB of random-access memory (RAM), and a Linux-based operating system provided by Google within the Colab Notebook environment. Statistical analysis was performed using R version 3.4.4, and machine learning tasks were executed using the scikit-learn framework.

4.1. (FA1): To what extent do the thirty hybrid architectures developed on Apache Spark demonstrate performance in classifying breast cancer?

As shown in Tables 3, 4, and 5 the four metrics results over the datasets. In the discussion of the hybrid architecture assessments across the break-his, FNAC, and ICIAR datasets, nuanced insights emerge regarding the influence of specific feature extractors and classifiers on the overall performance. For the break-his dataset, logistic regression-based hybrids consistently demonstrated superior accuracy, particularly when leveraging ResNet50 for feature extraction across varying magnification factors (MF). Noteworthy is the commendable performance of random forest-based architectures, excelling with ResNet50, while support vector machine-based hybrids showcased optimal accuracy with this particular feature extractor. Naive Bayes-based hybrids similarly favored ResNet50, outshining other extractors. On the contrary, the use of VGG19, VGG16, and Inception-V3 often resulted in comparatively lower accuracy scores, indicating the importance of the choice of feature extractor. Transitioning to the FNAC dataset, logistic regression-based hybrids achieved peak accuracy when employing ResNet50. However, divergent feature extractor influences were observed for other classifiers, underscoring the nuanced impact of the pairing on classification performance. In the case of the ICIAR dataset, logistic regression-based hybrids demonstrated optimal accuracy with ResNet50, emphasizing the significance of feature extraction technique selection. This detailed discussion highlights the intricate interplay between specific feature extractors and classifiers, shedding light on their distinctive roles in shaping the performance outcomes of hybrid architectures across diverse datasets. The findings underscore the importance of thoughtful selection to enhance classification accuracy in pathology image analysis.

4.2. (FA2): Is there a feature extractor approach that obviously exceeds others when used in a heterogeneous design?

This section aims to evaluate the impacts of five deep learning techniques as feature extractors on the performance of six machine learning classifiers. The objective is to identify the feature extraction approaches that significantly influence classification performance across the BreakHis, FNAC, and ICIAR datasets. The Scott Knott statistical test was employed, utilizing accuracy scores from heterogeneous architectures for each classifier, to group strategies with comparable prediction abilities, regardless of the feature extraction method. Figures 2 and 3 present the results of the Scott Knott test based on accuracy, highlighting the highest-performing CNN approaches independent of the classifiers.

For the magnification factor (MF) 40X, the SK test revealed that ResNet50 outperformed other CNN techniques, followed by Inception-V3, VGG19, Xception, with VGG16 being the least performant. Similarly, for MF 100X, ResNet50 was the top-performing CNN technique, followed by Xception, Inception-V3, VGG19, and VGG16 as the least performant. For MF 200X, ResNet50 led, followed by Xception, Inception-V3, VGG16, with VGG19 being the least performant. In the case of MF 400X, ResNet50 was the most effective, followed by VGG19, Inception-V3, VGG16, and Xception as the least performant.

Table 3. The accuracy of the thirty-heterogenous architecture over the breakHis dataset

Classifier-ML	Features extraction	400X	200X	100X	40X
Logistic regression	Inception-V3	89.91%	91.71%	90.92%	93.42%
	Resnet50	94.84%	96.64%	96.59%	98.20%
	VGG16	89.05%	91.20%	90.41%	93.02%
	VGG19	89.27%	89.34%	89.60%	91.63%
	Xception	89.27%	92.30%	90.35%	93.62%
Random forest	Inception-V3	81.57%	82.64%	80.94%	84.23%
	Resnet50	89.44%	86.98%	84.71%	88.29%
	VGG16	83.02%	76.98%	75.37%	79.65%
	VGG19	83.39%	83.42%	82.06%	89.41%
	Xception	80.74%	85.28%	82.07%	83.02%
SVM	Inception-V3	89.31%	87.07%	87.71%	93.41%
	Resnet50	89.91%	90.18%	91.10%	92.76%
	VGG16	89.41%	88.39%	91.10%	89.70%
	VGG19	88.76%	83.97%	82.09%	88.69%
	Xception	88.65%	89.06%	90%	90.51%
GBT	Inception-V3	73.59%	73.82%	75.55%	77.08%
	Resnet50	85.50%	86.15%	88.41%	84.21%
	VGG16	76.73%	82.08%	76.86%	80.33%
	VGG19	78.12%	77.12%	79.58%	76.77%
	Xception	71.91%	75.20%	73.75%	77.26%
Naïve Bayes	Inception-V3	80.60%	83.01%	80.47%	86.46%
	Resnet50	79.25%	80.37%	80.75%	79.12%
	VGG16	76.32%	77.61%	73.17%	77.81%
	VGG19	79.04%	78.11%	76.53%	79.22%
	Xception	79.66%	82.45%	81.57%	80.79%
Decision tree	Inception-V3	72.04%	74.71%	73.20%	75.43%
	Resnet50	87.37%	84.52%	86.03%	83.87%
	VGG16	77.84%	77.16%	76.03%	81.19%
	VGG19	78.26%	78.11%	77.54%	76.19%
	Xception	75.15%	79.62%	74.52%	77.35%

Table 4. The thirty-hybrid architecture's accuracy on the FNAC-dataset

Classifier-ML	Features extraction	Precision	Recall	F1-Score	Accuracy
Logistic regression	Inception-V3	92.22%	92.19%	92.18%	92.19%
	Resnet50	95.31%	95.31%	95.31%	95.31%
	VGG16	85.97%	85.93%	85.93%	85.93%
	VGG19	85.93%	85.93%	85.93%	85.93%
	Xception	94.54%	94.53%	94.53%	94.53%
Random forest	Inception-V3	89.10%	89.06%	89.06%	89.06%
	Resnet50	82.18%	80.29%	80.26%	80.29%
	VGG16	76.75%	75%	74.65%	75%
	VGG19	78.61%	78.12%	78.06%	78.12%
	Xception	78.98%	78.91%	78.90%	78.91%
SVM	Inception-V3	89.06%	89.06%	89.06%	89.06%
	Resnet50	82.04%	81.75%	81.79%	81.75%
	VGG16	82.44%	82.03%	81.99%	82.03%
	VGG19	79.83%	78.90%	78.77%	78.90%
	Xception	89.07%	88.28%	88.23%	88.28%
GBT	Inception-V3	78.98%	78.91%	78.90%	78.91%
	Resnet50	97.54%	79.56%	79.48%	79.56%
	VGG16	71.38%	67.96%	66.80%	67.96%
	VGG19	70.92%	68.75%	68.04%	68.75%
	Xception	73.17%	72.65%	72.55%	72.65%
Naïve Bayes	Inception-V3	81.27%	81.25%	81.24%	81.25%
	Resnet50	69.57%	69.34%	68.70%	69.34%
	VGG16	74.25%	74.21%	74.19%	74.21%
	VGG19	78.21%	78.12%	78.09%	78.12%
	Xception	80.47%	80.47%	80.46%	80.47%
Decision tree	Inception-V3	78.98%	78.90%	78.90%	78.91%
	Resnet50	76.62%	75.78%	75.63%	75.78%
	VGG16	71.39%	67.97%	66.81%	67.97%
	VGG19	69.66%	67.18%	66.28%	67.18%
	Xception	73.17%	72.65%	72.55%	72.65%

Table 5. The thirty-hybrid architecture's accuracy on the ICIAR dataset

Classifier-ML	Features extraction	Precision	Recall	F1-Score	Accuracy
Logistic regression	Inception-V3	89.99%	89.70%	89.67%	89.70%
	Resnet50	97.21%	97.05%	97.05%	97.05%
	VGG16	88.23%	88.23%	88.23%	88.23%
	VGG19	87.50%	87.50%	87.49%	87.50%
	Xception	86.03%	86.02%	86.02%	86.02%
Random forest	Inception-V3	80.51%	80.16%	79.98%	80.16%
	Resnet50	81.39%	81.39%	81.39%	81.39%
	VGG16	76.51%	76.47%	76.47%	76.47%
	VGG19	70.71%	70.58%	70.58%	70.58%
	Xception	82.49%	82.35%	82.35%	82.35%
SVM	Inception-V3	91.73%	91.73%	91.73%	91.73%
	Resnet50	89.90%	89.92%	89.90%	89.92%
	VGG16	87.58%	87.50%	87.50%	87.50%
	VGG19	86.86%	86.76%	86.74%	86.76%
	Xception	87.23%	86.76%	86.69%	86.76%
GBT	Inception-V3	65.45%	65.44%	65.35%	65.44%
	Resnet50	75.78%	73.64%	73.84%	73.64%
	VGG16	67.70%	67.64%	67.54%	67.64%
	VGG19	71.50%	71.32%	71.31%	71.32%
	Xception	73.17%	72.65%	72.55%	72.65%
Naïve Bayes	Inception-V3	84.22%	83.47%	83.25%	83.47%
	Resnet50	89.17%	89.14%	89.07%	89.14%
	VGG16	82.50%	81.81%	81.57%	81.81%
	VGG19	86.03%	86.02%	86.02%	86.02%
	Xception	83.82%	83.86%	83.82%	83.82%
Decision tree	Inception-V3	66.87%	66.94%	66.89%	66.94%
	Resnet50	71.63%	71.32%	71.29%	71.32%
	VGG16	62.51%	62.50%	62.50%	62.50%
	VGG19	71.50%	71.32%	71.31%	71.32%
	Xception	68.23%	67.64%	67.19%	67.64%

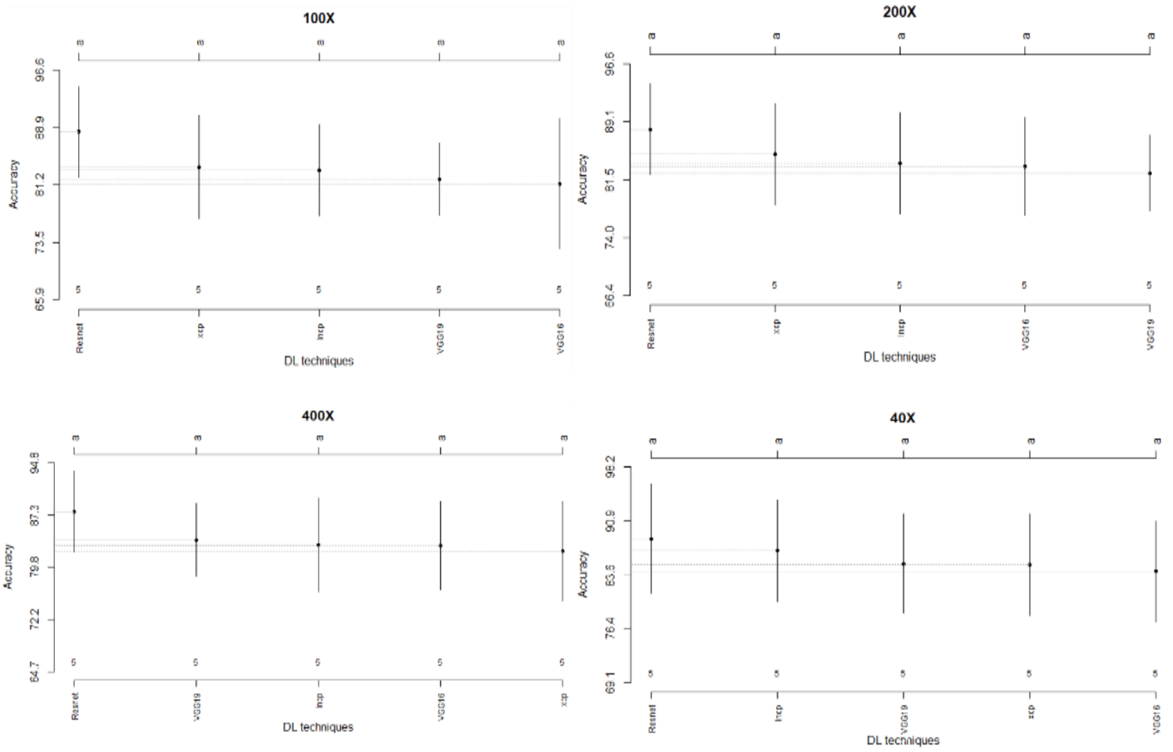

Figure 2. The outcomes of the Scott Knott test conducted on the five DL techniques using the BreakHis dataset

In the FNAC dataset, ResNet50 exhibited the highest performance, followed by Xception, VGG19, and VGG16, while Inception-V3 performed the least effectively. Similarly, within the ICIAR dataset, ResNet50 once more demonstrated superior performance among the CNN techniques, followed by Xception, VGG19, Inception-V3, and VGG16, with the latter being the least performant. Overall, ResNet50 consistently achieved the best results as a feature extractor across all classifiers and datasets, while VGG16 and Inception-V3 were identified as the less performant CNN techniques in specific magnification factors and datasets, respectively. The SK test on the three datasets collectively identified Inception-V3 as the least performing CNN technique globally.

4.3. (FA3): Is there a hybrid design that obviously outperforms others, irrespective of the selection of feature extractor and classifier?

In this analysis, we conduct a thorough evaluation of the predictive performance showcased by the top heterogeneous models from six classifiers across multiple datasets. Employing the Scott Knott (SK) test with accuracy as the benchmark, we delve into the rankings of diverse architectures over the BreakHis dataset. Figure 3 visually presents the SK statistical test results, highlighting the superior heterogeneous design that outperforms its counterparts, regardless of the feature extractor and classifier used. Figure 4 comprehensively depict the SK test outcomes for the six classifiers (LR, SVM, RF, NB, DT, and GBT) using features extracted with five architectures (VGG16, Inception-V3, Xception, VGG19, ResNet50) across different magnification factors and datasets (BreakHis, ICIAR, and FNAC). These figures unveil distinct groupings and performances. For the BreakHis dataset, the SK test identifies logistic regression with ResNet50 (RESLR) as the top-performing architecture and decision tree with Inception-V3 (DTINCP) as the less-performing one. In the FNAC dataset, nine classes emerge, with RESLR standing out as the best architecture and DTV19 as the least-performing. The ICIAR dataset reveals seven classes, with RESLR once again asserting itself as the superior architecture and DTV16 as the less-performing one. In summary, RESLR consistently emerges as the best architecture across all three datasets, while DTV16, DTV19, and DTINCP are identified as less-performing architectures in ICIAR, FNAC, and BreakHis datasets, respectively. This discussion underscores the nuanced interplay of classifiers and feature extractors, providing valuable insights for optimizing performance in pathology image analysis applications.

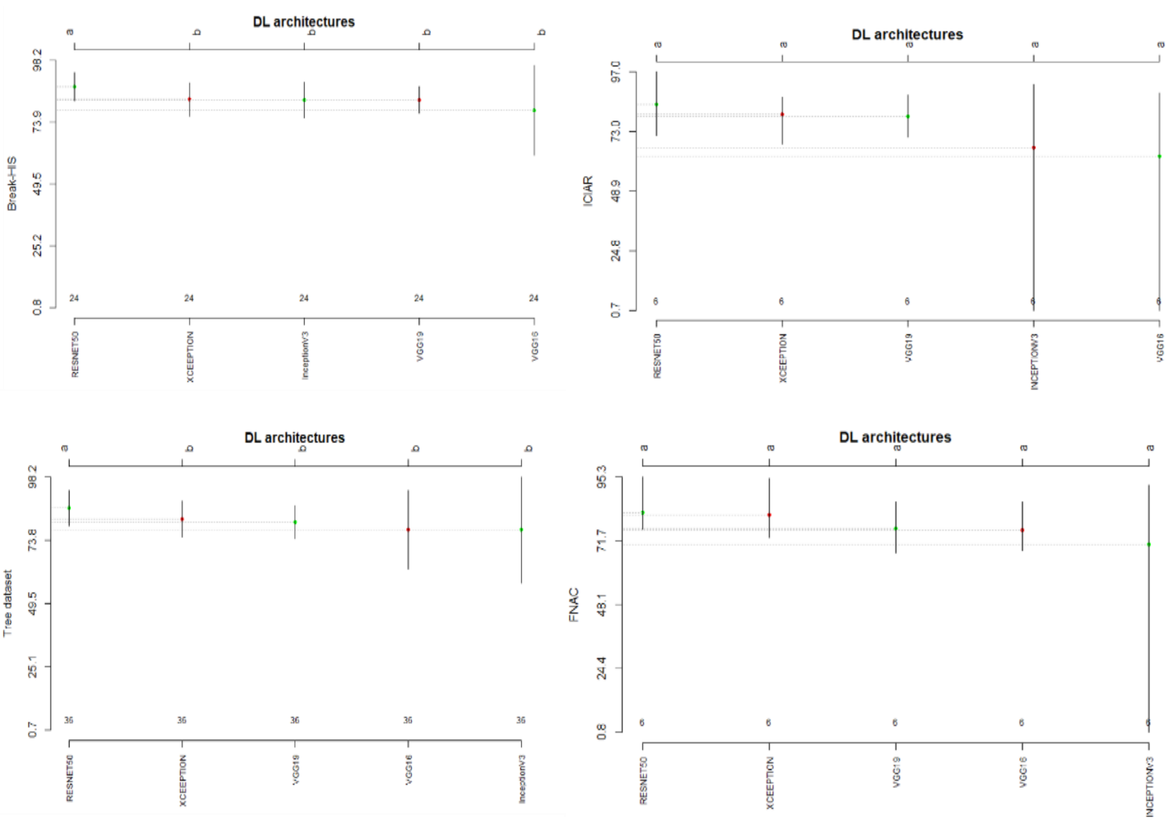

Figure 3. the outcomes of the Scott Knott test based on the five deep learning-techniques over the Three datasets

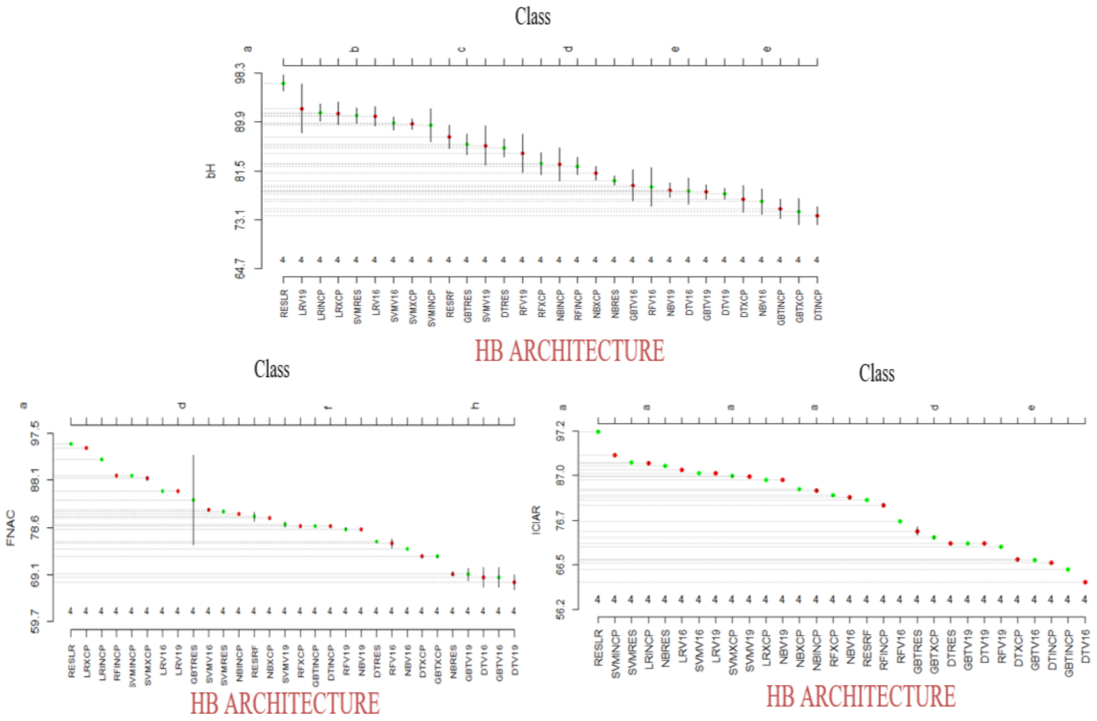

Figure 4. The outcomes of the Scott Knott test conducted on the heterogenous architectures over the Three datasets

4.4. (FA4): Is there a combination design that clearly outperforms others, independent of the dataset?

The hybrid architecture, utilizing logistic regression for classification and ResNet50 as a feature extractor, demonstrated strong performance across multiple datasets. Specifically, on the BreakHis dataset, the RESLR architecture achieved impressive accuracy rates of 98.20%, 96.59%, 96.64%, and 94.84% for MF categories (40, 100, 200, and 400), respectively. Furthermore, on the FNAC dataset, RESLR exhibited notable accuracy, reaching 95.31%. Similarly, on the ICIAR dataset, the same architecture showcased excellent performance with an accuracy of 97.05%. In summary, the RESLR hybrid architecture consistently delivered outstanding results across diverse datasets, including BreakHis, FNAC, and ICIAR.

4.5. (FA5): Did the implementation of Apache Spark have a noticeable impact on the outcomes, indicating an improvement?

Table 6 illustrates the training and testing times of diverse hybrid architectures on the FNAC dataset, highlighting distinct performance characteristics. Notably, the ResNet50 with logistic regression combination stands out for its efficient training (278.7/s) and swift testing (107.225/s). Conversely, ResNet50 with SVM exhibits longer training (2680.073/s) but compensates with notably faster testing (76.414/s). Inception-V3 with GBT demonstrates the highest training speed (5979.279/s) and remarkably fast testing (101.24/s). VGG16 with RF achieves the fastest training (1955.498/s), while VGG19 with SVM shows the longest training (11642.771/s). These findings underscore trade-offs in training and testing efficiency among hybrid architectures, offering insights for optimal combinations. In summary, the ResNet50 and logistic regression hybrid proves most time-efficient during training, while random forest with VGG16 excels in testing. Our work presents superior times compared to existing studies, enabling faster predictions with Apache Spark's integration, enhancing large dataset utilization.

Our study underscores the importance of specific feature extractors and classifiers in shaping breast cancer classification performance across diverse datasets. Logistic regression-based hybrid architectures, especially those incorporating ResNet50, consistently exhibited superior accuracy, highlighting the importance of strategic component selection in pathology image analysis.

Among deep learning techniques, ResNet50 consistently outperformed other architectures across various magnification factors and datasets. Notably, logistic regression with ResNet50 (RESLR) emerged as the top-performing architecture, demonstrating its superiority across the Break-His, FNAC, and ICIAR datasets.

The integration of Apache Spark yielded notable improvements in training and testing times. The ResNet50 and logistic regression hybrid proved the most time-efficient during training, while random forest with VGG16 excelled in testing. These findings offer valuable insights into optimal design considerations and computational efficiency, paving the way for advancements in breast cancer classification through hybrid architectures.

RESLR, excels with impressive accuracy: 98.20% on break-his, 97.05% on ICIAR, and 95.31% on FNAC. Comparatively to traditional deep learning VGG16 achieves 94% accuracy, while ResNet50 follows closely with 91.7% [12]. Traditional machine learning models like KNN excel at 97.51%, surpassing NB, which achieves a respectable 96.19% [13]. This comprehensive evaluation underscores the varied strengths of combined models in the context of medical imaging tasks.

Table 6. The time spent training and testing on the FNAC dataset

HB architecture FNAC	Training time /s	Testing time/s
GBT RES	214.261	110.999
NB RES	250.304	146.290
RF RES	898.961	227.285
LR RES	278.7	107.225
SVM RES	2680.073	76.414
DT RES	510.999	75.805
DT INCPV3	1332.621	158.363
LR INCPV3	1549.278	149.053
NB INPV3	274.485	784.541
RF INCPV3	1846.066	320.538
GBT INCPV3	5979.279	101.24
SVM INCPV3	5745.457	152.736
LR VGG19	853.057	296.951
GBT VGG19	863.172	305.187
NB VGG19	607.915	373.673
DT VGG19	1606.117	206.219
RF VGG19	2418.767	390.701
SVM VGG19	11642.771	284.268
GBT VGG16	812.729	237.743
DT VGG16	1552.262	197.547
RF VGG16	1955.498	31.015
LR VGG16	780.311	242.881
NB VGG16	494.266	304.940
SVM VGG16	9730.273	242.412
NB XCP	358.489	232.330
LR XCP	3080.16	204.076
DT XCP	2250.009	199.409
RF XCP	2375.203	265.488
SVM XCP	6957.896	195.474
GBT XCP	18318.06	200.734

5. CONCLUSION

This study advances breast cancer imaging classification by exploring 30 hybrid architectures, emphasizing diverse designs and effective feature extraction with classifier combinations. Notably, the RESLR architecture, utilizing ResNet50 as a feature extractor and logistic regression as a classifier, achieves outstanding accuracy of 98.20%, 96.59%, 96.64%, and 94.84% across various magnifications on the Break-His dataset. Additionally, RESLR demonstrates high accuracy on the ICIAR (97.05%) and FNAC (95.31%) datasets, promising enhanced diagnostic capabilities in pathology image analysis. Introducing Apache Spark, our study showcases its transformative impact on handling extensive datasets, improving both training and testing phases. This efficiency has the potential to revolutionize computational pathology. Our commitment extends to refining existing architectures and developing novel combinations to enhance diagnostic accuracy. Furthermore, we plan to investigate the scalability and adaptability of these architectures in real-world clinical settings, ensuring their practical utility and impact on patient care.

REFERENCES

[1] WHOQOL Group, "The World Health Organization quality of life assessment (WHOQOL): position paper from the World Health Organization," *Social Science and Medicine*, vol. 41, no. 10, pp. 1403–1409, Nov. 1995, doi: 10.1016/0277-9536(95)00112-K.

[2] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, Feb. 2021, doi: 10.3322/caac.21660.

[3] H. Zerouaoui and A. Idri, “Reviewing machine learning and image processing based decision-making systems for breast cancer imaging,” *Journal of Medical Systems*, vol. 45, no. 1, Jan. 2021, doi: 10.1007/s10916-020-01689-1.

[4] T. Chaymae, H. Elkhatir, and A. Otman, “Recent advances in machine learning and deep learning in vehicular ad-hoc networks: a comparative study,” in *The Proceedings of the International Conference on Electrical Systems & Automation*, Springer Singapore, 2022, pp. 1–14.

[5] C. Taib, O. Abdoun, and E. Haimoudi, “Performance evaluation of diagnostic and classification systems using deep learning on Apache Spark,” in *Lecture Notes in Mechanical Engineering*, Springer International Publishing, 2023, pp. 145–154.

[6] C. Taib, El. L. Haimoudi, and O. Abdoun, “Pneumonia classification using hybrid architectures based on ensemble techniques and deep learning,” in *Lecture Notes in Networks and Systems*, vol. 772 LNNS, Springer Nature Switzerland, 2023, pp. 389–399.

[7] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, and S. Pandiyan, “Skin disease detection and segmentation using dynamic graph cut algorithm and classification through naive Bayes classifier,” *Measurement: Journal of the International Measurement Confederation*, vol. 163, Oct. 2020, doi: 10.1016/j.measurement.2020.107922.

[8] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[9] M. A. Aswathy and M. Jagannath, “Detection of breast cancer on digital histopathology images: Present status and future possibilities,” *Informatics in Medicine Unlocked*, vol. 8, pp. 74–79, 2017, doi: 10.1016/j.imu.2016.11.001.

[10] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157, doi: 10.1109/iccv.1999.790410.

[11] A. Khandakar *et al.*, “A machine learning model for early detection of diabetic foot using thermogram images,” *Computers in Biology and Medicine*, vol. 137, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104838.

[12] N. S. Ismail and C. Sovuthy, “Breast cancer detection based on deep learning technique,” in *2019 International UNIMAS STEM 12th Engineering Conference, EnCon 2019 - Proceedings*, Aug. 2019, pp. 89–92, doi: 10.1109/EnCon.2019.8861256.

[13] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, “Breast cancer classification using machine learning,” in *2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT)*, Apr. 2018, pp. 1–4, doi: 10.1109/EBBT.2018.8391453.

[14] O. M. Al-hazaimeh, A. A. Abu-Ein, N. M. Tahat, M. A. Al-Smadi, and M. M. Al-Nawashi, “Combining artificial intelligence and image processing for diagnosing diabetic retinopathy in retinal fundus images,” *International Journal of Online and Biomedical Engineering*, vol. 18, no. 13, pp. 131–151, Oct. 2022, doi: 10.3991/ijoe.v18i13.33985.

[15] H. Zerouaoui, A. Idri, F. Z. Nakach, and R. El Hadri, “Breast fine needle cytological classification using deep hybrid architectures,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12950, Springer International Publishing, 2021, pp. 186–202.

[16] F. R. Cordeiro, W. P. Santos, and A. G. Silva-Filho, “A semi-supervised fuzzy GrowCut algorithm to segment and classify regions of interest of mammographic images,” *Expert Systems with Applications*, vol. 65, pp. 116–126, Dec. 2016, doi: 10.1016/j.eswa.2016.08.016.

[17] Zhengyou Zhang, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron,” in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 454–459, doi: 10.1109/AFGR.1998.670990.

[18] S. Guan and M. Loew, “Breast cancer detection using transfer learning in convolutional neural networks,” in *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Oct. 2017, pp. 1–8, doi: 10.1109/AIPR.2017.8457948.

[19] H. Zerouaoui and A. Idri, “Deep hybrid architectures for binary classification of medical breast cancer images,” *Biomedical Signal Processing and Control*, vol. 71, Jan. 2022, doi: 10.1016/j.bspc.2021.103226.

[20] F. O. Ozkok and M. Celik, “A hybrid CNN-LSTM model for high resolution melting curve classification,” *Biomedical Signal Processing and Control*, vol. 71, Jan. 2022, doi: 10.1016/j.bspc.2021.103168.

[21] M. Sewak, P. Vaidya, C.-C. Chan, and Zhong-Hui Duan, “SVM approach to breast cancer classification,” in *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, Aug. 2008, pp. 32–37, doi: 10.1109/imsccs.2007.46.

[22] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, “A hybrid algorithm for lung cancer classification using SVM and Neural Networks,” *ICT Express*, vol. 7, no. 3, pp. 335–341, Sep. 2021, doi: 10.1016/j.icte.2020.06.007.

[23] K. Batko and A. Ślęzak, “The use of big data analytics in healthcare,” *Journal of Big Data*, vol. 9, no. 1, Jan. 2022, doi: 10.1186/s40537-021-00553-4.

[24] H. Kadkhodaei, A. M. Eftekhari Moghadam, and M. Dehghan, “Big data classification using heterogeneous ensemble classifiers in Apache Spark based on MapReduce paradigm,” *Expert Systems with Applications*, vol. 183, Nov. 2021, doi: 10.1016/j.eswa.2021.115369.

[25] L. Li, J. Lin, Y. Ouyang, and X. (Robert) Luo, “Evaluating the impact of big data analytics usage on the decision-making quality of organizations,” *Technological Forecasting and Social Change*, vol. 175, Feb. 2022, doi: 10.1016/j.techfore.2021.121355.

[26] G. Aresta *et al.*, “BACH: grand challenge on breast cancer histology images,” *Medical Image Analysis*, vol. 56, pp. 122–139, Aug. 2019, doi: 10.1016/j.media.2019.05.010.

[27] A. R. Saikia, K. Bora, L. B. Mahanta, and A. K. Das, “Comparative assessment of CNN architectures for classification of breast FNAC images,” *Tissue and Cell*, vol. 57, pp. 8–14, Apr. 2019, doi: 10.1016/j.tice.2019.02.001.

[28] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A dataset for breast cancer histopathological image classification,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.

[29] S. Karen and Z. Andrew, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2015.

[30] K. J. Worsley, “A non-parametric extension of a cluster analysis method by Scott and Knott,” *Biometrics*, vol. 33, no. 3, Sep. 1977, doi: 10.2307/2529369.

BIOGRAPHIES OF AUTHORS

Chaymae Taib a Ph.D. student at the Computer Science Department of Abdelmalek Essadi University, has a Master's degree in big data and IoT from ENSAM Casablanca, which she earned in 2020. Her ongoing research focuses on neural networks, machine learning, data preprocessing, big data, and intelligent systems. With a collection of published works, she can be reached via email: taib.chaymae@etu.uae.ac.ma.

Otman Abdoun is currently a professor in the Department of Computer Science at the Faculty of science, Abdelmalek Essaadi University, Tetouan, Morocco. He earned his Ph.D. degree in computer science from Ibn Tofail Kenitra University in Morocco. His primary research interests lie in multi-agent systems, optimization algorithms, and E-learning. With numerous publications to his credit. He can be reached via email: abdoun.otman@gmail.com.

Elkhatir Haimoudi having successfully earned his Ph.D. degree in computer science from Ukraine University, El Khatir Haimoudi now serves as a professor in the Computer Science Department of the Polydisciplinary Faculty at Larache, Abdelmalek Essadi University. His passion for research encompasses areas such as neural networks, deep learning, pattern recognition, and classification problems. With an array of publications to his credit. He can be reached via email: helkhatir@gmail.com.