

## Brain tumor detection in the Spark system

Soumia Benkrama, Nour Elhouda Hemdani

Laboratory of TIT, Department of Computer Science, Faculty of Exact Sciences, Tahri Mohammed University, Bechar, Algeria

### Article Info

#### Article history:

Received Nov 14, 2022

Revised Mar 19, 2023

Accepted Apr 2, 2023

#### Keywords:

Brain tumors  
Convolutional neural network  
Deep learning  
EfficientNetB1  
Spark system

### ABSTRACT

Machine learning (ML) and computer vision systems revolutionized the world, especially deep learning (DL) for convolutional neural networks, which has proven breakthroughs in brain tumor (BT) diagnosis. This study investigates a convolutional neural network (CNN) approach for image classification for BT detection using the EfficientNetB1 architecture with global average pooling (GAP) layers in a big data setting. A classification layer is done with a softMax layer. The system is created in the Apache Spark environment. Spark system is a unified and ultra-fast analysis engine for large-scale data processing. It is mainly dedicated to big data and deep learning (DL). Experiments are carried out using the brain magnetic resonance imaging (MRI) dataset containing 3,264 MRI scans to predict the performance of the model. The dataset is decomposed into two datasets. The model's performance was assessed and compared to existing models, it yielded a high precision, precision, and f1-score. In our work, we have achieved an accuracy of 97% and a performance of 98% on a dataset of 3,064 brain MRI images.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Soumia Benkrama  
Laboratory of TIT, Department of Computer Sciences, Faculty of Exact Sciences  
Tahri Mohammed University  
Independence Road B.P 417, Bechar 08000, Algeria  
Email: benkrama.soumia@univ-bechar.dz

## 1. INTRODUCTION

Brain cancer is a cell development in the cerebrum, which can be harmless (non-dangerous) or threatening (carcinogenic). It may be having a cerebral origin or have invaded the brain after developing in another region of the body (metastasis). The fusion of medical imaging and artificial intelligence allows us to diagnose brain cancers more accurately and quickly than before. Automatic identification of a tumor as early as possible is essential for patient survival.

Recently, important advances have been made in this area using machine learning methods and specialized models [1]. Deep learning (DL) and classification constitute computational techniques [2]. Deep learning aims to recognize features. Lower level functions help generate higher-level functions. Deep learning has been proven to identify high-level features and give better results in classification [3]. Several works using DL for the detection of BT can be found in [4]-[7].

This work combines image processing and big data classification and brain tumor (BT) detection. We use the Spark system on Python to perform big data analysis of BT images. The dataset is supplied to the computing server system DataFrame from pyspark. Throughout the experimental phase, we used a dataset of magnetic resonance imaging (MRI) scans of different types of BT. This paper makes the following main contributions:

- Provides a computational framework for the classification of BT of MRI image slices in a big data environment.
- Application of complex neural networks in model data models (unstructured).

- Expose the high accuracy of the dataset in brain tumor detection.

This paper proposes the creation of a convolutional neural network (CNN) model that classifies brain tumors and learns the model using data (MRI of the brain) in Spark system. The main advantage of Spark is its speed. Spark was designed end-to-end with performance in mind, using in-memory computing and other optimizations to achieve this. It is also known for its ease of use and sophisticated analytics. Indeed, it provides simple application programming interfaces (APIs) for working with large datasets. We will therefore present a series of relevant work that relates to our topic and the proposed approach and ultimately to experimental results.

## 2. RELATED WORK

In neuro-oncology, tumors are often examined using MRI. The detection of tumor lesions is a critical step in examining tumors from images. Due to the popularity of deep learning methods, several research papers have emerged due to the benefits they provide in terms of data processing power and automation. Several DL detection architectures are currently proposed to automatically classify BT [8], [9]. The goal of DL is to process large amounts of data and also automatically identify and extract features without human intervention. Big data is distinguished by its volume, and it is also known for its heterogeneity of formats and structures, as well as its need for processing speed. The following sections discuss related works to classify and identify brain tumors.

Cheng *et al.* [8] proposed a model to increase the classification performance of BT by improving the tumor field. They retrieved several features to increase the performance of the model, including a bag of words and a grayscale co-occurrence matrix. To classify the features, they used the support vector machine (SVM). The model was 91.28% accurate on a dataset of 3,064 MR images of the brain.

Paul *et al.* [10] proposed a classification of BT based on CNN. The model obtained an accuracy of 90.26%. The model shows that diminishing image size may ameliorate driving performance in training. Tahir *et al.* [11] proposed a model based on MRI scans to increase classification performance. This model extracted the basic features of Daubechies wavelets. The model attained an accuracy of 86% on the data containing 3,064 brain MRI images using SVM. Ghassemi *et al.* [12] suggested a multi-class model based on a DNN for the classification of BT. The algorithm utilizes data augmentation techniques to extract data and form a neural network as discrimination within a generative antagonist network. To designate tumor classes the fully connected (FC) layers were replaced. The model achieved an accuracy of 95.6% on the random split and 93.01% on the inserted split. Palanisamy and Thirunavukarasu [13] used a pretrained GoogLeNet to extract attributes from brain MRIs for the classification of BT and utilize deep transfer learning (DTL) to fine-tune item classification. They performed 5-fold cross-validation on the MRI datasets and an accuracy of 98%. Das *et al.* [14] proposed the classification of BT based on a CNN. To pre-process the MRI images. They applied Gaussian filters and histogram equalization methods with a detection accuracy of 94.39%.

Choudhury *et al.* [15] proposed a computational system involving CNNs to classify MRI-based BT images into two classes: cancer and no cancer. The execution rate of this method is 96.08%. Pathak *et al.* [16] also used CNN models for the classification and segmentation of BT. A CNN model was applied to divide the input images into cancer and no cancer to segment the tumor afterward. Shafi *et al.* [17] suggested an ensemble learning approach to group lesions of cancers and autoimmune diseases using MRI of patients with BT. The method describes how pretreatment functions are extracted and classified. The preprocessing stage uses tumor and lesion regions of interest (ROI), Collwet normalization, and Lloyd max quantification. The experimental results indicate that the training and test accuracies of the suggested model are 97.957% and 97.744% respectively. Mehrotra *et al.* [18]. suggested a model that increases the accuracy of brain MRI image recognition by using transfer learning methods. They used different models such as SqueezeNet, ResNet50, ResNet101 GoogLeNet, and AlexNet. This pretrained, the final characteristics are classified with the Softmax layer. The process begins with a dataset of MR brain images that have been collected and classified into the tumor and no tumor. The proposition includes the following steps: preprocessing, data division, extraction of features, and tumor type classification. The configured arrangement achieves performance with 99,04% accuracy. Vankdothu *et al.* [19]. suggested a system for the detection of BT. This system is included different categories such as image, feature extraction, preprocessing, and image classification. They used adaptive filters in a preprocessing step to remove noise from MRI images. For extracting features, they used deep learning models to classify image types and used recurrent convolutional neural networks (RCNN) for the classification process. This classifier achieved 95.17% accuracy.

Shinde *et al.* [20], determined the analysis of different ML methods, such as SVM and logistic to classify tumors as benign and malignant, and discrete wavelet transforms to extract features from synthetic data available on web resources open access series of imaging studies (OASIS) and Alzheimer's disease neuroimaging initiative (ADNI). Although there are various techniques for detecting BT. The automatic classification of the diverse types of tumors remains essential. Therefore, this work targets the classification of BT established on DL in an Apache Spark environment.

### 3. PROPOSED APPROACH

This study combines big data concepts with image processing concepts. The goal of this work is to improve tumor detection strategies using brain MRI image classification. This article presents a CNN method for the autonomous classification of BT in big data using Apache Spark. Apache Spark is used for its speed in processing large amounts of data. It enables cluster machines to perform extensive analysis. It mainly focuses on big data, ML, and CNN. The goal is to achieve high classification accuracy.

Our system consists of classifying images using the CNN approach, which is a type of DL. First, we have an MRI image as input. With this image our system will be able to diagnose the state of the patient if he is healthy or sick in the case where the patient is sick, we detect the class of tumor. The general architecture of the system used is presented in Figure 1. The aim is the early detection of BT which will allow professionals to advance more efficient treatment regimens to lead to healthier outcomes for patients [21].

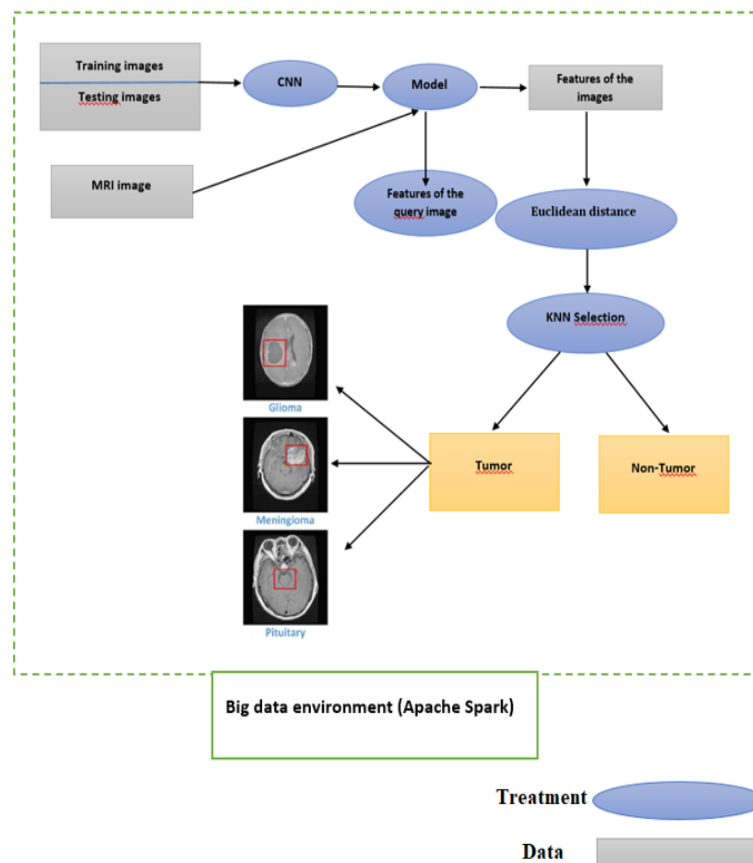


Figure 1. The general architecture

Before proceeding to classification, we start by using CNN on our data (training and testing). CNN runs this with its convolutional layers with various filter sizes, which makes the network more visible [22]. CNN models usually contain convolutional fully connected, pooling, and classification layers [23]. In our work, we use the EfficientNetb1 CNN model, which is extensively used in literature, we use it for its high accuracy, good efficiency, and less time and with adjusted standards. The EfficientNetB1 model increases accuracy while diminishing the number of parameters [24]. In EfficientNetb1 they increased the mesh width and resolution equally. This architecture used Bottleneck convolution [25]. Softmax is used in its final layer, the classification layer.

The EfficientNetb1 model is used for multi-scale and multinetwork network structures. We use the EfficientNetb1 library by this structure in jupyter:from tensorflow.keras.applications import EfficientNetB1. Thus, we will have obtained deep features of the images from the EfficientNetb1 model. In this study, we used the Epoch variable to experiment. The epoch corresponds to training on all the data, and the greater this number greater the accuracy should be obtained.

If there is an MRI image as input, this query image (the image is inserted by the user to find out which class belongs) goes directly to the model to obtain the characteristics of this image. After, we calculate the euclidean distance between the base of the characteristics and the characteristics of the image query. After we apply the k-nearest neighbors (KNN). KNN algorithm used in the classification. Handles training data and categorizes new test data established on distance measurements. It finds the k-NN to the test data to then make the selection to diagnose the image of the patient either healthy or sick. If the result is sick, the type of tumor is detected.

## 4. EXPERIMENTAL RESULTS AND DISCUSSIONS

### 4.1. Dataset

The publicly available Kaggle dataset is used in our work with a total of 3,264 MRI images, decomposed into test and training images of 394 and 2,870 images respectively [26]. We compared our results with other classifiers to make sure that the model used was properly implemented. MRI images are associated with BT, such as glioma, meningioma, pituitary, and no cancer. The informations of the dataset are presented in Table 1. The four MRI classes are described in Figure 2, where Figure 2(a) illustrated the class of images not containing cancer, and Figures 2(b)-(d) illustrated different tumors of the brain. The total number of images is 3,264.

Table 1. The brain image base

	Training	Testing
No cancer	395	105
Gliomas	826	100
Meningiomas	822	115
Pituitary	827	74
Total	2870	394

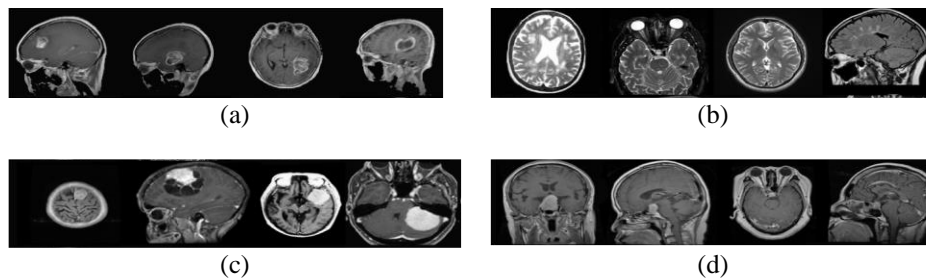


Figure 2. Dataset of brain MRI images: (a) no-cancers, (b) gliomas, (c) meningiomas, and (d) pituitary

### 4.2. Evaluation metrics used

In this study, we used several evaluations. The objective of these measures is the evaluation of the performance rate of our system. System performance is tested and confirmed; the precision of the system, recall, f1-score, avg-macro, weighted avg, and accuracy were determined with the quantification of the predicted classes according to the following quantities: the number of false negatives (FN), the number of false positives (FP), the number of true negatives (TN) and the number of true positive (TP). The mathematical representation is defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$f1 - score = 2 \times \frac{recall \times precision}{recall + precision} \quad (4)$$

$$avg - macro = average\ of\ each\ one\ (precision, recall, f1\ score)$$

$$weighted\ avg = \frac{(P_0 \times C_0) + (P_1 \times C_1) + (P_2 \times C_2) + (P_3 \times C_3)}{(C_0 + C_1 + C_2 + C_3)} \tag{5}$$

where  $C \sum$  test values (Table 1) and  $P$  is the precision of each one of the classes.

### 4.3. Experimental test and results

To evaluate our system, we beginning by installing Apache Spark under jupyter, we use the following command in the Anaconda prompt: Pip install pyspark. For access to jupyter through the big data environment, we use the following command: Pyspark. This command opens the jupyter window automatically. Table 2 illustrates the parameters used in the system proposed. The results obtained are presented in Table 3 which presents the comparison of the accuracy of the system over different numbers of epochs.

Table 2. The parameters used

Model and methods	Parameters
Algorithm	EfficientNetB1
Epoch numbers	Used as a variable (experiments)
Batch size	32 images
Pooling layer	Avg pooling
Activation function	SoftMax
Environment and language	Pyspark 'Big Data', Python (Anaconda Jupyter)

Table 3. Comparison of accuracy on different epochs numbers

Epoch	Accuracy
05	0.89
10	0.96
15	<b>0.98</b>
20	<b>0.98</b>

From epoch 15 there is a stabilization in accuracy versus variable epoch. In the rest of this work, we use epoch=20, and accuracy=0.97. Comparison training and validation (accuracy/loss) by epochs are presented in Figure 3. Then, we have the confusion matrix (error matrix) in Figure 4. It represents the percentage of the prediction of the true images in the diagonal square and the other for the false prediction. For example, in a square (1,1)=0.98 (98%) and (0,1)=0.02 (2%). The percentage of true prediction is 98% and of the false prediction is 2%. Other evaluation measures are in Figure 5. Where 0, 1, 2, 3: glioma, no\_cancer (normal), meningioma and pituitary.

Epochs vs. Training and Validation Accuracy/Loss

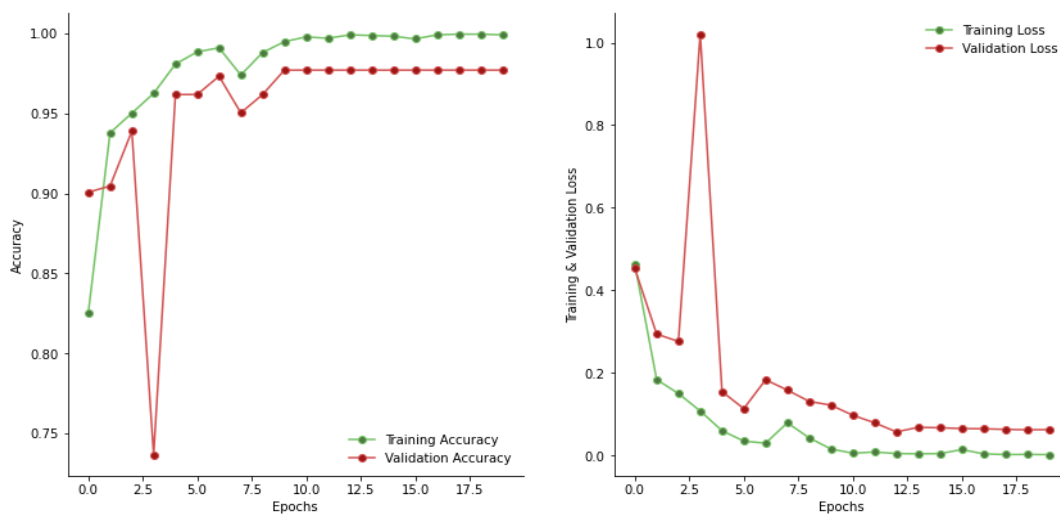


Figure 3. Comparison training and validation (accuracy/loss) by epochs

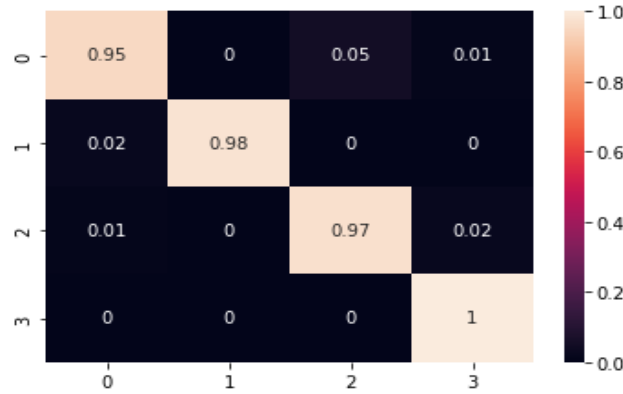


Figure 4. Confusion matrix for 04 class classification on MRI dataset

	precision	recall	f1-score	support
0	0.97	0.95	0.96	168
1	1.00	0.98	0.99	108
2	0.96	0.97	0.96	201
3	0.97	1.00	0.99	176
accuracy			0.97	653
macro avg	0.98	0.97	0.97	653
weighted avg	0.97	0.97	0.97	653

Figure 5. Results of the evaluation measures used

The study of Vankdothu and Hameed [19] divided their research into several parts, including: pre-processing, segmentation, and MRI image feature extraction. BT images were segmented using improved clustering of K-means. The approach achieved an accuracy of 95.17%. Aamir *et al.* [27] divided the dataset into two groups 70% (training) and 30% (validation). In our study, if the dataset for training increases the performance will increase thereafter. In Table 4, we analyze the performance of our system with studies conducted with the same dataset [26].

Table 4. Comparison of works with proposed approach using the same dataset

Writer	Method and model	Performance
Vankdothu and Hameed [19]	Model CNN- long short-term memory (LSTM)	Accuracy=92%. Precision=95,17 %. Recall=98%.
Sartaj [26].	Model CNN, EfficientNetB0, Resnet	Accuracy=98,95 %.
Proposed approach	Apache Spark, CNN, EfficientNetB1	Accuracy=98%. Recall=98,55%. Precision=98%.




### 5. CONCLUSION

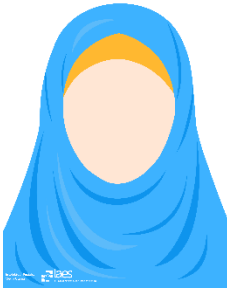
The aim of our study is to obtain good classification accuracy in the detection of BT. We implemented our method using "TensorFlow" and "Keras" in "Python" and "Apache Spark" because it is an efficient programming language for fast work. The system proposed will help doctors accurately identify brain tumors using MRI. The originality of this work is the use of efficientnetB1 in the Spark system. The training was faster and more accurate. Experiments show that efficientnetB1 provides 97% accuracy in Apache Spark. In conclusion, we found that the proposed strategy is applicable to classify various forms of BT. In the future, we will expand our dataset by adding more images as well as applying other pre-trained CNN algorithms to increase and compare results and choose the best model.




## REFERENCES

- [1] 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.
- [2] R. Thillaikkarasi and S. Saravanan, “An enhancement of deep learning algorithm for brain tumor segmentation using kernel based CNN with M-SVM,” *Journal of Medical Systems*, vol. 43, no. 4, 2019, doi: 10.1007/s10916-019-1223-7.
- [3] J. Amin, M. Sharif, N. Gul, M. Yasmin, and S. A. Shad, “Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network,” *Pattern Recognition Letters*, vol. 129, pp. 115–122, Jan. 2020, doi: 10.1016/j.patrec.2019.11.016.
- [4] M. L. Rahman, A. W. Reza, and S. I. Shabuj, “An internet of things-based automatic brain tumor detection system,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 214–222, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp214-222.
- [5] H. E. Hamdaoui *et al.*, “High precision brain tumor classification model based on deep transfer learning and stacking concepts,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 24, no. 1, p. 167, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp167-177.
- [6] M. Ahmed, F. Khalifa, H. E. Moustafa, G. A. Saleh, and E. AbdElhalim, “A deep learning based system for accurate diagnosis of brain tumors using T1-w MRI,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, p. 1192, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1192-1202.
- [7] M. W. Nadeem *et al.*, “Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges,” *Brain Sciences*, vol. 10, no. 2, p. 118, Feb. 2020, doi: 10.3390/brainsci10020118.
- [8] J. Cheng *et al.*, “Enhanced performance of brain tumor classification via tumor region augmentation and partition,” *PLoS ONE*, vol. 10, no. 10, p. e0140381, Oct. 2015, doi: 10.1371/journal.pone.0140381.
- [9] S. Deepak and P. M. Ameer, “Brain tumor classification using deep CNN features via transfer learning,” *Computers in Biology and Medicine*, vol. 111, p. 103345, Aug. 2019, doi: 10.1016/j.combiomed.2019.103345.
- [10] J. S. Paul, A. J. Plassard, B. A. Landman, and D. Fabbri, “Deep learning for brain tumor classification,” in *Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*, Mar. 2017, vol. 10137, p. 1013710, doi: 10.1117/12.2254195.
- [11] B. Tahir *et al.*, “Feature enhancement framework for brain tumor segmentation and classification,” *Microscopy Research and Technique*, vol. 82, no. 6, pp. 803–811, Jun. 2019, doi: 10.1002/jemt.23224.
- [12] N. Ghassemi, A. Shoeibi, and M. Rouhani, “Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images,” *Biomedical Signal Processing and Control*, vol. 57, p. 101678, Mar. 2020, doi: 10.1016/j.bspc.2019.101678.
- [13] V. Palanisamy and R. Thirunavukarasu, “Implications of big data analytics in developing healthcare frameworks – A review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 4, pp. 415–425, Oct. 2019, doi: 10.1016/j.jksuci.2017.12.007.
- [14] S. Das, O. F. M. R. R. Aranya, and N. N. Labiba, “Brain tumor classification using convolutional neural network,” in *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*, May 2019, vol. 68, no. 1, pp. 1–5, doi: 10.1109/ICASERT.2019.8934603.
- [15] C. L. Choudhury, C. Mahanty, R. Kumar, and B. K. Mishra, “Brain tumor detection and classification using convolutional neural network and deep neural network,” *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*, 2020, doi: 10.1109/ICCSEA49143.2020.9132874.
- [16] M. K. Pathak, M. Pavthawala, M. N. Patel, D. Malek, V. Shah, and B. Vaidya, “Classification of brain tumor using convolutional neural network,” in *Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019*, Jun. 2019, pp. 128–132, doi: 10.1109/ICECA.2019.8821931.
- [17] A. S. M. Shafi, M. B. Rahman, T. Anwar, R. S. Halder, and H. M. E. Kays, “Classification of brain tumors and auto-immune disease using ensemble learning,” *Informatics in Medicine Unlocked*, vol. 24, p. 100608, 2021, doi: 10.1016/j.imu.2021.100608.
- [18] R. Mehrotra, M. A. Ansari, R. Agrawal, and R. S. Anand, “A transfer learning approach for AI-based classification of brain tumors,” *Machine Learning with Applications*, vol. 2, p. 100003, Dec. 2020, doi: 10.1016/j.mlwa.2020.100003.
- [19] R. Vankdothu and M. A. Hameed, “Brain tumor MRI images identification and classification based on the recurrent convolutional neural network,” *Measurement: Sensors*, vol. 24, p. 100412, Dec. 2022, doi: 10.1016/j.measen.2022.100412.
- [20] A. S. Shinde, B. M. Mahendra, S. Nejakar, S. M. Herur, and N. Bhat, “Performance analysis of machine learning algorithm of detection and classification of brain tumor using computer vision,” *Advances in Engineering Software*, vol. 173, p. 103221, Nov. 2022, doi: 10.1016/j.advensoft.2022.103221.
- [21] Y. Guan *et al.*, “A framework for efficient brain tumor classification using MRI images,” *Mathematical Biosciences and Engineering*, vol. 18, no. 5, pp. 5790–5815, 2021, doi: 10.3934/MBE.2021292.
- [22] X. Wang and W. Zhang, “Anti-occlusion face recognition algorithm based on a deep convolutional neural network,” *Computers and Electrical Engineering*, vol. 96, p. 107461, Dec. 2021, doi: 10.1016/j.compeleceng.2021.107461.
- [23] M. Toğaçar, “Detection of segmented uterine cancer images by Hotspot Detection method using deep learning models, Pigeon-Inspired Optimization, types-based dominant activation selection approaches,” *Computers in Biology and Medicine*, vol. 136, p. 104659, Sep. 2021, doi: 10.1016/j.combiomed.2021.104659.
- [24] L. Gaur, U. Bhatia, N. Z. Jhanjhi, G. Muhammad, and M. Masud, “Medical image-based detection of COVID-19 using Deep Convolution Neural Networks,” *Multimedia Systems*, vol. 1, pp. 1–10, Apr. 2021, doi: 10.1007/s00530-021-00794-6.
- [25] P. Kaur, B. Singh, K. Bbsbec-Baba, B. S. Bahadur, A. Partap, and S. Pharwaha, “Evaluation of Base Networks for Object Classification and Detection,” *International Journal of Advanced Research in Engineering and Technology*, vol. 11, no. 12, pp. 3132–3141, 2020, doi: 10.34218/IJARET.11.12.2020.295.
- [26] S. Sartaj, “Brain Tumor Classification (MRI),” *kaggle*, 2023. [Online]. Available: <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset?datasetId=1608934&searchQuery=EFFI>.
- [27] M. Aamir *et al.*, “A deep learning approach for brain tumor classification using MRI images,” *Computers and Electrical Engineering*, vol. 101, p. 108105, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108105.

**BIOGRAPHIES OF AUTHORS**

**Dr. Soumia Benkrama**    is currently an associate professor at the faculty of exact sciences, computer science department, University Tahri Mohamed-Béchar, Algeria. She received a master's and Ph.D. degrees in computer science at the University of Science and Technology Mohamed Boudiaf, Oran, Algeria. Her research areas are image processing, machine learning, pattern recognition, and computational intelligence. She can be contacted at email: [benkrama.soumia@univ-bechar.dz](mailto:benkrama.soumia@univ-bechar.dz).



**Hemdani Nour Elhouda**    was born in Bechar in 1999. She had her baccalaureate in 2017. She studied computer science at the university of Tahri Mohammed in Bechar. She received a license diploma in 2020 and a master's in artificial intelligence in 2023. Her research interests are machine learning (deep learning), image processing, and computer vision. She can be contacted at email: [hamhouda7@gmail.com](mailto:hamhouda7@gmail.com).