Received: 25 August 2022 | Revised: 17 June 2023 | Accepted: 10 July 2023

DOI: 10.4218/etrij.2022-0088

ORIGINAL ARTICLE

ETRI Journal WILEY

Transfer-learning-based classification of pathological brain magnetic resonance images

Serkan Savaş[1] | **Çağrı Damar**[2]

[1]Department of Computer Engineering, Kırıkkale University, Kırıkkale, Turkey

[2]Department of Radiology, Gaziantep University, Gaziantep, Turkey

Correspondence
Serkan Savaş, Department of Computer Engineering, Kırıkkale University, 71450 Kırıkkale, Turkey.
Email: serkansavas@kku.edu.tr

Funding information
This research did not receive funding.

Abstract

Different diseases occur in the brain. For instance, hereditary and progressive diseases affect and degenerate the white matter. Although addressing, diagnosing, and treating complex abnormalities in the brain is challenging, different strategies have been presented with significant advances in medical research. With state-of-art developments in artificial intelligence, new techniques are being applied to brain magnetic resonance images. Deep learning has been recently used for the segmentation and classification of brain images. In this study, we classified normal and pathological brain images using pretrained deep models through transfer learning. The EfficientNet-B5 model reached the highest accuracy of 98.39% on real data, 91.96% on augmented data, and 100% on pathological data. To verify the reliability of the model, fivefold cross-validation and a two-tier cross-test were applied. The results suggest that the proposed method performs reasonably on the classification of brain magnetic resonance images.

KEYWORDS

brain MRI, deep learning, EfficientNet, transfer learning, two-tier cross-test, white matter

1 | INTRODUCTION

The brain can be considered as the most sensitive, important, and complex organ in the human body. Hence, the diagnosis and treatment of any brain disease is challenging for physicians as well as critical and vital for patients. In recent years, computer-aided treatment methods have attracted increasing interest in diagnosis, treatment, and surgery. Manual treatments are being gradually replaced by their computer-aided counterparts.

The sensitive structure of the brain is equally susceptible to many kinds of infections and other disorders, such as tumors, Alzheimer's disease, multiple sclerosis, alcoholism, drugs, amnesia, altitude sickness, autism, epilepsy, trauma, hemorrhage, stroke, and mental retardation. These conditions adversely affect the brain [1]. Brain disorders affect millions of people worldwide and causes a major disease burden. Many of these disorders are chronic and incurable, and most available treatment options are either inadequate or ineffective. Consequently, these disorders incur significant social and economic costs. Despite the urgent clinical need, minimal progress has been made in the development of new treatments for most common brain disorders, and many drugs currently prescribed are based on outdated science [2]. Therefore, brain disease diagnosis and examination must be performed using state-of-the-art technologies.

Magnetic resonance imaging (MRI) scans of the brain can show brain structures, shrinkage, vascular irregularities, and other structural changes that may indicate

cognitive dysfunction [3]. White-matter hyperintensities (WMHs) seen on T2-weighted or fluid-attenuated inversion recovery (FLAIR) MRI sequences correspond to brain regions affected by previous or ongoing conditions. These abnormal signal alterations can be focal, multifocal, or confluent and may even be located in more extensive areas of the white matter [4]. WMHs are observed in diverse circumstances, including inherited diseases such as leukodystrophies, cerebral blood supply disorders, myelin-damaging diseases, infections, autoimmune toxic metabolic factors, neoplasia, and traumatic injury [5, 6]. Age-related WMHs emerge due to a chronic cerebrovascular ischemia reflected as leukoaraiosis, and the accompanying signal changes in elderly patients increase the risk of stroke, cognitive impairment, dementia, and death [7–9]. In addition, a significant association has been found between WMHs and late-life depression and bipolar disorder [10, 11].

Along with parameters such as age, sex, genetic predisposition, environmental conditions, lifestyle, and risk factors (e.g.,, diabetes mellitus, cardiovascular diseases, hypercholesterolemia, hypertension, smoking, alcohol consumption, and drug addiction), clinical and laboratory features as well as brain MRI evaluation results help clinicians in diagnosis [7]. However, conventional brain MRI evaluation is observer dependent and time consuming. Progress in computer-aided radiologic imaging technologies and deep learning (DL) for neuroimaging has led to the rapid development of diagnostic and treatment strategies. Some studies have described the use of artificial intelligence (AI) for better patient stratification and evaluation of treatment responses in both clinical care and drug trials [12].

DL is a promising area of machine learning (ML), which has been developed as a solution to AI problems and has grown tremendously over the past decade. It mainly relies on the deep neural network (DNN) structure. DL is aimed to produce solutions based on learning datasets, including large datasets, rather than on a separate model specific to each study [13]. It is widely used in various fields such as image processing, computer identification, autonomous vehicles, natural language processing, and audio and video recognition [14]. DL approaches for automated medical decision making are also required in medical image processing because nonautomated processes are labor intensive, often expensive, and highly prone to human errors [13]. Therefore, ML and DL have become increasingly common in medicine, particularly over the past decade. These techniques have been used in different studies, from detecting Alzheimer's disease [15] to detecting the novel coronavirus disease [16], investigating the disease effects [17], early predicting type 2 diabetes [18], and diagnosing cancer [19]. Although medical image processing studies on brain MRI have been mainly focused on segmentation, classification has been explored but not extensively applied.

This study was aimed to isolate magnetic resonance (MR) images, including areas with WMHs, from normal individuals using pretrained DL models. Thus, an innovative solution was developed to support clinical decision making. To this end, we answered the following research questions (RQs):

- RQ1: Can pretrained DL models be used to suitably classify different medical images?
- RQ2: Does data augmentation contribute to learning?
- RQ3: What are the learning differences between real and synthetic images?
- RQ4: Does a two-tier cross-test contribute to classification robustness?

In addition, we improved the classification of brain MR images by using real data, achieving a high accuracy and avoiding overfitting, unlike similar studies (see Section 5).

The remainder of this paper is organized as follows. Section 2 presents a brief overview of recent studies on brain image segmentation and classification. Section 3 details the materials and methods of the study. Section 4 reports and analyses the test results, which are discussed in Section 5. Finally, Section 6 summarizes the contributions of this study and directions of future work.

2 | RELATED WORK

AI [20] has been applied to many fields including medicine, with applications such as diagnosis, treatment, imaging, data security, and drug delivery. ML, a subfield of AI, has been intensively used for diagnosis and treatment based on medical imaging over the past two decades. In particular, DL, a subfield of ML, has been used since the 2010s because of its increasing performance [21].

Studies on brain MRI are extensive regarding segmentation and classification. For instance, some morphological and ML techniques have been applied to brain MRI. With state-of-the-art developments in AI, new techniques are being applied to brain MR images. Recently, DL models and techniques have been used for brain image segmentation. An artificial neural network was presented and optimized for automatically segmenting white-matter changes in MR images to produce probabilistic maps [22]. Because convolutional neural networks (CNNs) are powerful for image processing, they are widely applied to MR images. A CNN was proposed for

segmentation of hyperintensities and distinguishing between WMHs and stroke lesions using a UResNet architecture [23]. Arı and Hanbay [24] developed a regional CNN for automated tumor detection using T1- and T2-weighted MR images. A CNN was also combined with random forest regressors to predict the expanded disability status of patients with multiple sclerosis at 2 years solely based on age, sex, and FLAIR MRI data [25]. It was also combined with SLIC0 clustering for candidate segmentation of white-matter lesions [26].

Beyond the detection of white-matter diseases, abnormalities in brain images have been detected using classification techniques. Saritha and others [27] proposed an approach integrating wavelet entropy-based spider web plots, and a probabilistic neural network was developed for classifying brain MR images as normal or indicating stroke, infectious disease, degenerative disease, or brain tumor. Lu and others [28] used wavelet entropy as a feature and a kernel-based extreme learning machine (ELM) as the classifier. Lu and others [29] introduced an approach to automatically detect pathological brains in MRI based on a DL structure and transfer learning using AlexNet [30] for a pretrained network. Lu et at. [31] applied ResNet [32] and randomized neural networks to brain MR images. ResNet was the feature extractor with three neural networks: Schmidt network, random vector functional-link net, and ELM. The weights and biases in the three networks were trained using the chaotic bat algorithm.

In DL, transfer learning for classifying brain abnormalities is under development and becoming increasingly common. Existing studies have various limitations (see Section 5), including overfitting and lack of cross-validation. In this study, we obtained high accuracy and reliability by applying a two-tier cross-test to real data.

3 | MATERIALS AND METHODS

3.1 | Data and training

To obtain sufficient brain MR images, a retrospective evaluation was performed using the picture archiving and communication system of Gaziantep University, Faculty of Medicine, and the Department of Radiology. Brain MRI was performed using a 3-T magnet system (Philips Ingenia 3.0 T; Philips Medical Systems, the Netherlands). To test the developed CNN, 619 MR images of 619 patients (age range, 10–40 years) were examined for possible brain pathologies. Axial midline images of FLAIR sequences (repetition time/inversion time, 11 000/2800 ms; echo time, 120 ms; matrix size, 240 × 139, field of view, 230 × 189 × 144 mm; recon voxel size, 0.57 × 0.57 × 4.00 mm; slice thickness, 4 mm; number of slices, 32; signal-to-noise ratio, 1; number of signal averages/number of excitations, 1) including genu and splenium of the corpus callosum, lateral ventricles, foramen of Monro, bilateral basal ganglia, and thalamus were stored as 1196 × 1196-pixel images in the JPG format without any identity information. Images with insufficient quality in terms of spatial resolution or containing artifacts were discarded. The quality of acquired MR images can vary according to factors such as settings, operator characteristics, patient characteristics, artifacts, and file format, as described below.

- *Settings and operator characteristics*: There are several image acquisition parameters in the control panels of MRI systems. Radiology technicians can adjust the parameters to acquire MRI sequences. The signal-to-noise ratio, number of excitations, field of view, matrix size, and slice thickness are important factors whose changes can affect the image spatial resolution. During acquisition of brain MRI sequences, some technicians can use custom settings or modify the standard settings of the parameters to accelerate imaging.
- *Patient characteristics and artifacts*: The quality of the obtained MR images decreases when the patient moves in the gantry of the magnet during acquisition. In addition, specific artifacts can disrupt the image quality in MRI examinations.
- *File format*: Normally, the image format of MRI is DICOM (Digital Imaging and Communications in Medicine), and the files are converted from DICOM to JPEG. This conversion negatively affects the image quality.

From the dataset, we excluded low-quality DICOM images acquired with nonoptimal settings or artifacts that made diagnosis difficult and JPEG images with resolutions below 1196 × 1196 pixels.

Institutional Ethics Committee approval (granted on October 27, 2021, under number 722) was obtained for the study protocol. Images were classified as normal or pathological by a radiologist with at least 8 years of experience. Patients with headaches or other neurological complaints were physically examined by physicians. In these cases, blood tests and radiological examinations might be required. Radiologists recognized pathological changes or signals on MR images and reported these findings along with a preliminary diagnosis to physicians. There were different abnormal signals in the MRI sequences, which were depicted in grayscale. Although some major abnormalities could be easily defined, others, especially small ones, could be missed. This depended on the concentration and experience of the radiologist.

Therefore, the human dependence on MRI reporting was disadvantageous. Sample classes of images obtained from patients within the scope of ethical approval are shown in Figure 1. The abnormal white-matter signals are marked with arrows in Figure 1B.

The Kaggle [33] and Colab [34] platforms were used to train DL models on brain MR images. Both platforms allow researchers, students, and users from different fields to work online using powerful computational graphics processors on topics such as ML, data science, and data analysis. In this study, model training was performed using the Python programming language with the TensorFlow and Keras application libraries. After the models were trained, their graphs and weights were recorded and used for testing. Accuracy metrics of the models were visualized, and a confusion matrix was obtained and interpreted to evaluate their performance.

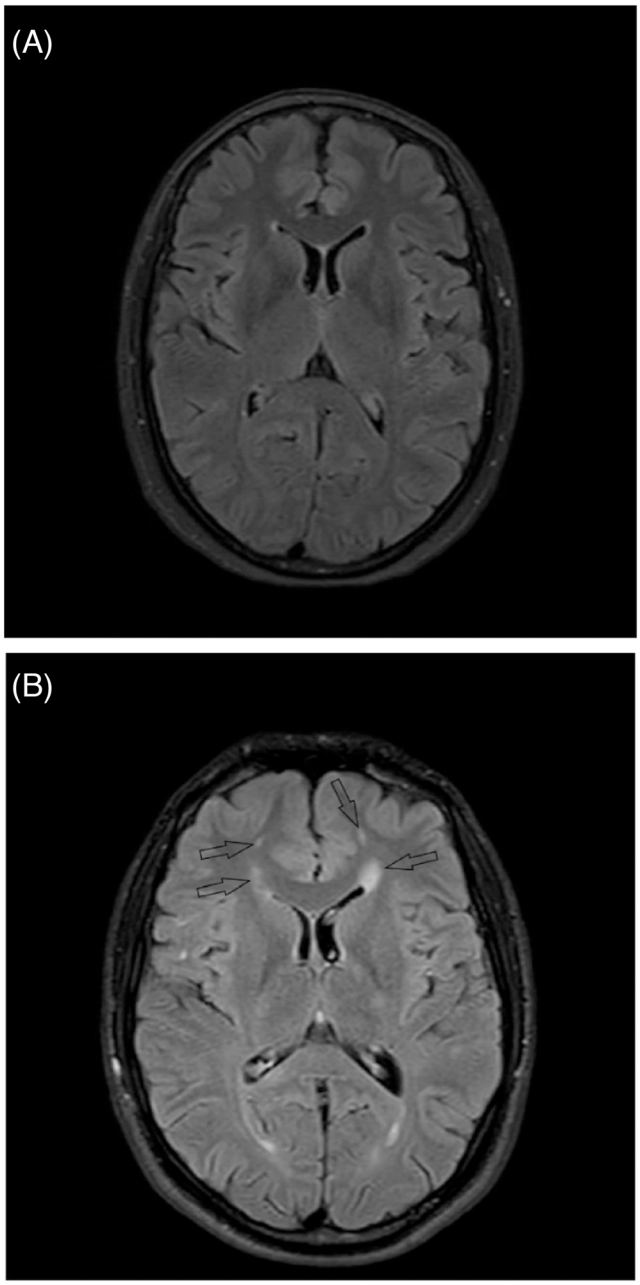


FIGURE 1 Examples of (A) normal and (B) abnormal brain magnetic resonance (MR) images in fluid-attenuated inversion recovery (FLAIR) sequence.

3.2 | Data augmentation

Unlike general ML, DL provides automatic feature extraction. DL requires feeding the model with excess data, which are important for successful training. As the amount of data increases, a model exhibits improved consistency and performance. On the other hand, the calculation costs and training time also increase, but these increases can be ignored or reduced using different solutions to automate learning.

Data acquisition is a major problem in biomedical imaging. The amount of available data is scarce, particularly for sensitive diseases. Additionally, owing to ethical issues, private data are not often encountered in databases worldwide. Researchers face various problems in medical studies, such as obtaining data for each study, obtaining ethical permission to gather the data, obtaining permission from patients to collect the data, interpreting these data by experts, and synchronizing data acquisition with the application time of the study to adapt to state-of-the-art solutions. Data augmentation can help overcoming data scarcity. This artificial data generation technique applies combinations of operations to existing data. Thus, data availability can be increased by adding synthetic data reflecting different variations. For images, data augmentation is typically achieved by adding various filters, flipping horizontally and vertically, rotating at different angles, scaling outward or inward, cropping random image sections, translating to x/y coordinates, and adding noise.

In this study, the images were rotated at different angles for data augmentation. First, 10% of the total images (62) were separated from the dataset for use in the two-tier cross-test validation. This is important for highlighting the robustness of the model and verifying the consistency of results. The original data were used to test the accuracy obtained after model training. Second, the remaining 557 images were rotated by 45° to synthesize new images, obtaining 5040 brain MR images for processing. Table 1 summarizes the statistics of the augmented data.

In most previous studies, owing to data scarcity, all data were included in the models, and the training accuracy was interpreted using these data. Various operations

were performed on the data to ensure the accuracy and reliability of the results. Similarly, we reserved 10% of the real patient data for testing the model (two-tier cross-test evaluation) with the weights recorded after training, validation, and testing. After applying data augmentation to the remaining 90% of real data, 20% of the data were reserved for initial testing after training and validation of the models. The remaining 80% of the available augmented data was divided into training (80%) and validation (20%) sets. During model training on the MR images, learning was primarily performed on the data (total training data) at the specific rates (80%), and the training result was validated (total validation data) over a certain rate of data (20%) in each epoch. The weights obtained after performing these operations over all the epochs to reach the optimum result were tested as the training result on an image set (total test data) (20%), which was unseen by the model during training and kept separately. All the validation samples listed in Table 1 were used to validate the accuracy of the results after each training epoch. The test samples, on the other hand, were stored for testing on other images that were not seen during training using the weights obtained after all epochs were completed and training finished. A block diagram of this study is shown in Figure 2.

The abovementioned data splitting is important in terms of two-tier cross-test validity and reliability after training and validation. The validation rates obtained in the first stage and the results obtained from both the augmented and real data on unseen datasets during testing

TABLE 1 Statistics of augmented data.

Image type	Normal	Pathological	
No. MR images	439	180	
No. two-tier cross-test MR images	44	18	Unseen test samples for two-tier cross-test
No. augmented samples	3600	1440	
No. training samples	2304	922	
No. validation samples	576	230	
No. test samples	720	288	Unseen augmented test samples

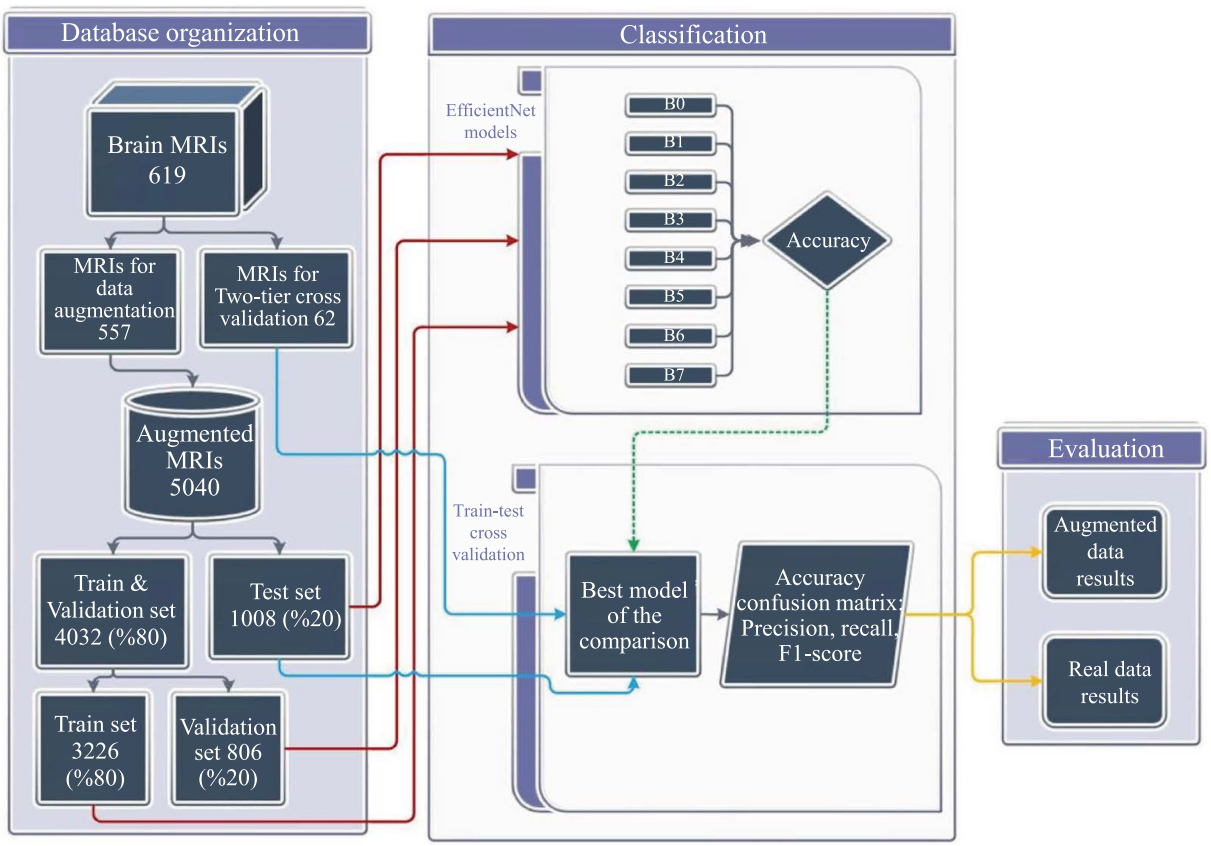


FIGURE 2 Block diagram of study.

were analyzed in this study. We conducted a two-tier cross-test. The generalizability of the models is particularly important for ML and DL models. The two-tier cross-test allows to obtain the accuracy of models from different studies after training by testing on real and synthetic data. Thus, the robustness and generalizability of the study can be ensured.

3.3 | Transfer learning

Since 1995, research on transfer learning has attracted attention in various areas such as knowledge transfer, lifelong learning, learning to learn, inductive transfer, multitask learning, knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, meta-learning, and incremental (cumulative) learning [35]. Over the past decade, with the development of DNNs, studies on transfer learning have become increasingly common. Transfer learning allows to solve problems involving small amounts of data [36]. This is achieved by retraining a previously successful DNN on a new dataset with corresponding changes in the hyperparameters and adjustments in the fully connected layer. The general structure of the network and feature extraction structure are preserved. Thus, a model used to solve a problem with time and effort savings can be reused to produce solutions for different problems. In recent years, studies on transfer learning have been conducted with high performance on different datasets.

3.4 | Model

Although deep architectures emerged earlier, the success of a deep CNN called AlexNet in the ImageNet competition held in 2012 increased the importance of DL. Subsequently, new and deeper models (i.e., DNNs) were introduced. In a DNN, two or more layers of hidden neural networks form a comprehensive connected architecture, ranging from simple to complex, with each layer trying to relate to the previous one. Owing to this comprehensiveness, inputs are examined in greater depth and detail, and features are extracted toward obtaining accurate results.

Several models to efficiently solve different applications using DL have been proposed over the past decade. Nevertheless, DL allows to generalize a solution to address various problems. Hence, pretrained models have been increasingly used in recent years [21]. In these models, solutions are obtained either by directly using the weights of the pretrained models or by updating these weights with additional layers using transfer learning. A well-known DL algorithm used in image classification is the CNN [13], like the one shown in Figure 3.

Our analysis was based on the EfficientNet model proposed by Tan and Le [37] to define a baseline for the classification of normal and pathological brain MR images. Using EfficientNet, we propose a compound scaling method that uniformly scales the network width, depth, and resolution with a set of fixed scaling coefficients. For example, if $2N$ more computational resources are used, the network can be expanded in depth by αN, width by βN, and image size by γN, where α, β, and γ are constant coefficients determined by a grid search on the original small model. Because model scaling does not change layer operator $\widehat{F}_i$ in the baseline, a good baseline should be determined [37]. EfficientNet comprises eight different models, from version B0 to B7. The baseline architecture of EfficientNet for version B0 is detailed in Table 2.

In Table 2, each row corresponds to stage i with $\widehat{L}_i$ layers, input resolution $\widehat{H}_i \times \widehat{W}_i$, and $\widehat{C}_i$ output channels. Starting from the baseline EfficientNet-B0, compound scaling is applied using (1) and (2) in two steps. Initially, $\varphi = 1$ to find the best values for α, β, and γ. Second, α, β, and γ are fixed as constants, and the baseline

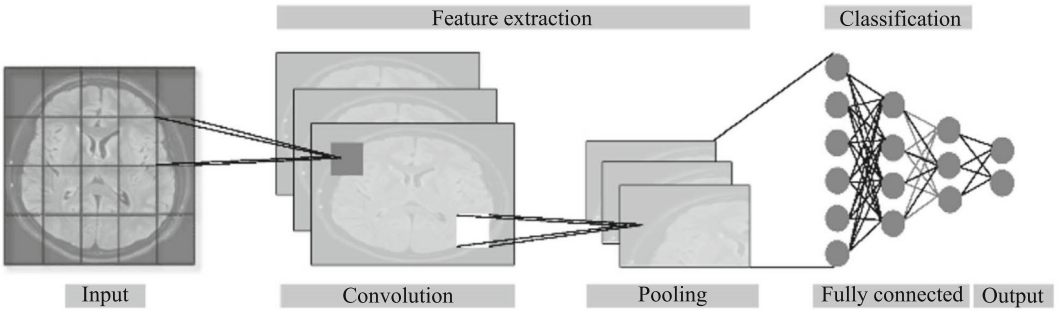


FIGURE 3 Sample convolutional neural network architecture.

network is scaled with different φ values to obtain EfficientNet-B1–B7 [37].

$$\begin{aligned} &\max_{d,w,r} \text{Accuracy}(N(d, w, r)) \\ &\text{s.t.} N(d, w, r) = \bigodot_{i=1,\ldots,s} \widehat{F}_i^{d\widehat{L}_i}\left(X_{\left\langle r.\widehat{H}_i, r.\widehat{W}_i, w.\widehat{C}_i\right\rangle}\right) \\ &\text{Memory}(N) \le \text{target\_memory} \\ &\text{FLOPS}(N) \le \text{target\_ flops}, \end{aligned} \tag{1}$$

where w, d, and r are coefficients for scaling the network width, depth, and resolution, respectively, and $\widehat{F}_i$, $\widehat{L}_i$, $\widehat{H}_i$, $\widehat{W}_i$, and $\widehat{C}_i$ are predefined parameters in the baseline network (see Table 2).

TABLE 2 EfficientNet-B0 baseline [37].

Stage i	Operator $\widehat{F}_i$	Resolution $\widehat{H}_i \times \widehat{W}_i$	No. channels $\widehat{C}_i$	No. layers $\widehat{L}_i$
1	Conv3 × 3	224 × 224	32	1
2	MBConv1, $k3 \times 3$	112 × 112	16	1
3	MBConv6, $k3 \times 3$	112 × 112	24	2
4	MBConv6, $k5 \times 5$	56 × 56	40	2
5	MBConv6, $k3 \times 3$	28 × 28	80	3
6	MBConv6, $k5 \times 5$	14 × 14	112	3
7	MBConv6, $k5 \times 5$	14 × 14	192	4
8	MBConv6, $k3 \times 3$	7 × 7	320	1
9	Conv1 × 1 & Pooling & FC	7 × 7	1280	1

Source: Reprinted with permission from Tan and Le [37], Copyright 2019. M. Tan and Q. Le.

$$\begin{aligned} &\text{depth}: d = \alpha^{\phi} \\ &\text{width}: w = \beta^{\phi} \\ &\text{resolution}: r = \gamma^{\phi} \\ &\text{s. t.} \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ &\alpha \ge 1,\ \beta \ge 1,\ \gamma \ge 1, \end{aligned} \tag{2}$$

where $\alpha, \beta,$ and γ are constants that can be determined by a small grid search. Coefficient ϕ is specified by the user and controls the additional resources available for model scaling, while $\alpha, \beta,$ and γ specify the assignment of the additional resources to the network width, depth, and resolution, respectively [37]. As with any DNN, the first EfficientNet structure is the root followed by architectural related structures common to its eight versions and final layers. The components of EfficientNet are shown in Figure 4 [38]. Each structure comprises seven blocks with increasing numbers of subblocks from EfficientNet-B0 to EfficientNet-B7 [38]. The modules and subblocks are shown in Figure 5.

Module 1 is used as a starting point for the subblocks, and Module 2 is used as a starting point for the first subblock of all the seven main blocks, except for the first one. Module 3 is a skip connection to all the subblocks, and Module 4 is used to combine the skip connections in the first subblock. Each subblock is connected to its preceding one via a skip connection and combined using Module 5. These modules are further combined to form subblocks that are used in a certain manner within the blocks. Subblock 1 is used only as the first subblock in the first block, and Subblock 2 is used as the first subblock in all the other blocks. Finally, Subblock 3 is used for any subblock, except for the first one in all blocks [38].

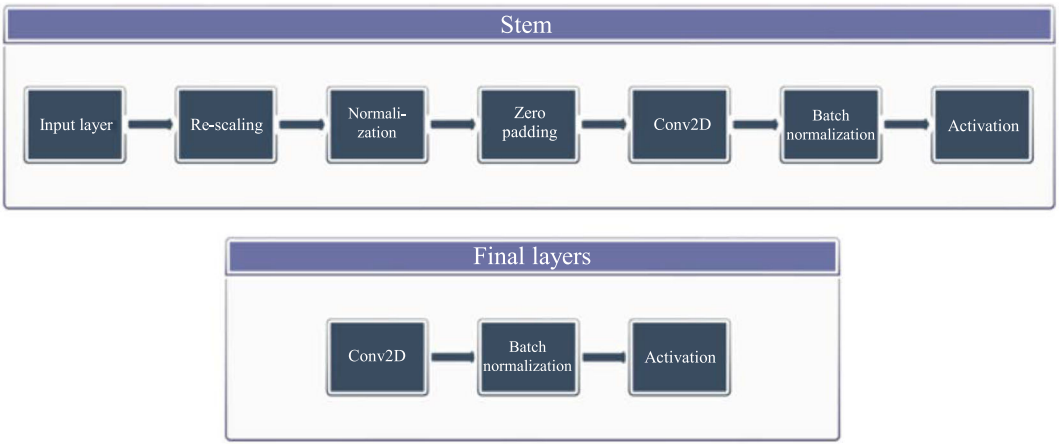


FIGURE 4 Root and final layers of EfficientNet [38] (reprinted with permission from V. Agarwal [38], copyright 2020 V. Agarwal).

These models were the most successful in classifying the stages of Alzheimer's disease on brain MR images in [39]. In this study, the EfficientNet-B0 model achieved an accuracy of 92.98%. In addition, EfficientNet-B3 (89.78%), EfficientNet-B2 (94.42%), and EfficientNet-B3 (97.28%) showed high precision, recall, and specificity (shown in parentheses), respectively. For EfficientNet versions B0–B7, we first determined the highest performing ones on augmented data. In other words, we performed testing to select the version to use before applying EfficientNet to the available data. All the models were subjected to training, validation, and testing using five-fold cross-validation. Table 3 lists the selection results.

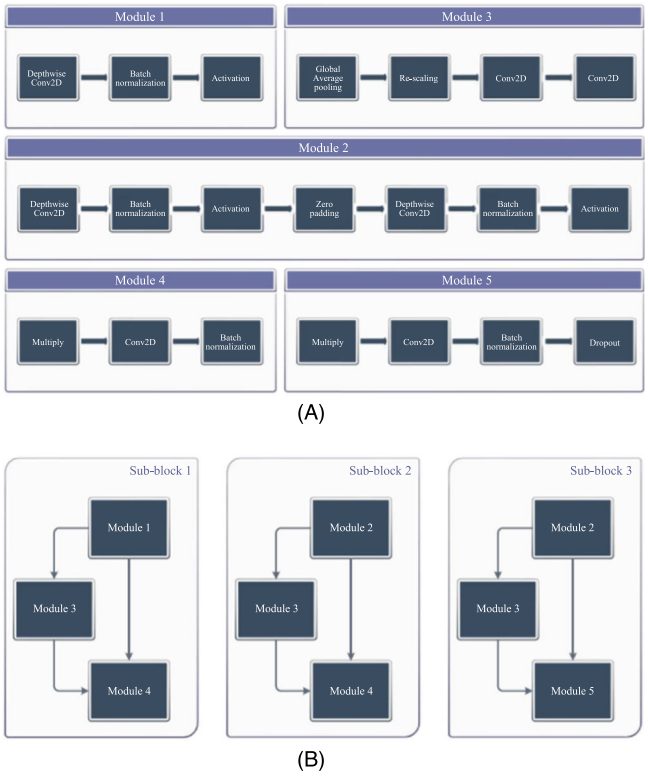


FIGURE 5 (A) Five modules used in EfficientNet and (B) subblocks [38] (reprinted with permission from V. Agarwal [38], copyright 2020 V. Agarwal).

Table 3 shows that the mean accuracy for EfficientNet-B5 was 91.29% and the standard deviation was 0.46% for fivefold cross-validation training. In addition, regarding the complexity matrix results, EfficientNet-B5 outperformed the other models in terms of precision, recall (sensitivity), specificity, and F1-score. Moreover, the recall of all the models was lower than the precision. During the two-tier cross-test, the evaluation metrics were obtained from real data. These results support the use of EfficientNet-B5 for transfer learning. The architecture of EfficientNet-B5 is shown in Figure 6.

To apply EfficientNet-B5, the fully connected layers were reduced to two classes for the classification of normal and pathological MR images. For reduction, two 4096 dense layers, one 1000 dense layer, and one 512 dense layer, were used. Dropout was applied between each dense layer (value of 0.5). The hyperparameters used in the model were specified as follows. The images were stored in the JPG format, and the input image size was 224×224 pixels. The number of epochs for the model was 250, and the batch size was set to 10. In DL, training is frequently conducted over 50–100 epochs. In this study, more training epochs were performed, obtaining 250 as the cutoff point for learning convergence. The models did not show considerable learning improvement after 250 epochs, which were determined for all the evaluated models. Consequently, training took longer than usual, especially considering health studies. Nevertheless, any improvement is important, thus justifying the longer training time. Adam optimization was applied with a learning rate of 0.00001. Binary cross-entropy was used as the loss function with accuracy as its metric. After repeated tests, the optimal parameter values were defined and used in the trained model.

TABLE 3 Test results of EfficientNet models.

Model	Test 1	Test 2	Test 3	Test 4	Test 5	Mean	*SD*	Precision	Recall (sensitivity)	Specificity	F1-score
EfficientNet-B0	86.90	86.41	87.00	87.10	86.51	86.78	0.31	81.25	72.22	93.33	76.43
EfficientNet-B1	85.42	87.00	87.40	86.81	86.90	86.71	0.75	80.91	61.81	94.12	70.02
EfficientNet-B2	87.50	87.00	88.10	88.19	86.20	87.40	0.83	82.35	69.86	94.02	75.59
EfficientNet-B3	89.08	89.19	88.00	88.99	88.78	88.81	0.48	88.12	61.81	96.62	72.64
EfficientNet-B4	89.68	88.69	87.80	89.08	89.29	88.91	0.72	92.86	59.06	97.21	72.13
EfficientNet-B5	91.17	90.67	91.27	91.36	91.96	91.29	0.46	93.17	76.33	97.76	84.04
EfficientNet-B6	89.98	87.60	87.50	88.89	87.59	88.31	1.10	89.05	62.15	96.90	72.93
EfficientNet-B7	89.48	90.77	90.27	89.38	90.11	90.00	0.58	91.59	70.83	97.53	79.88

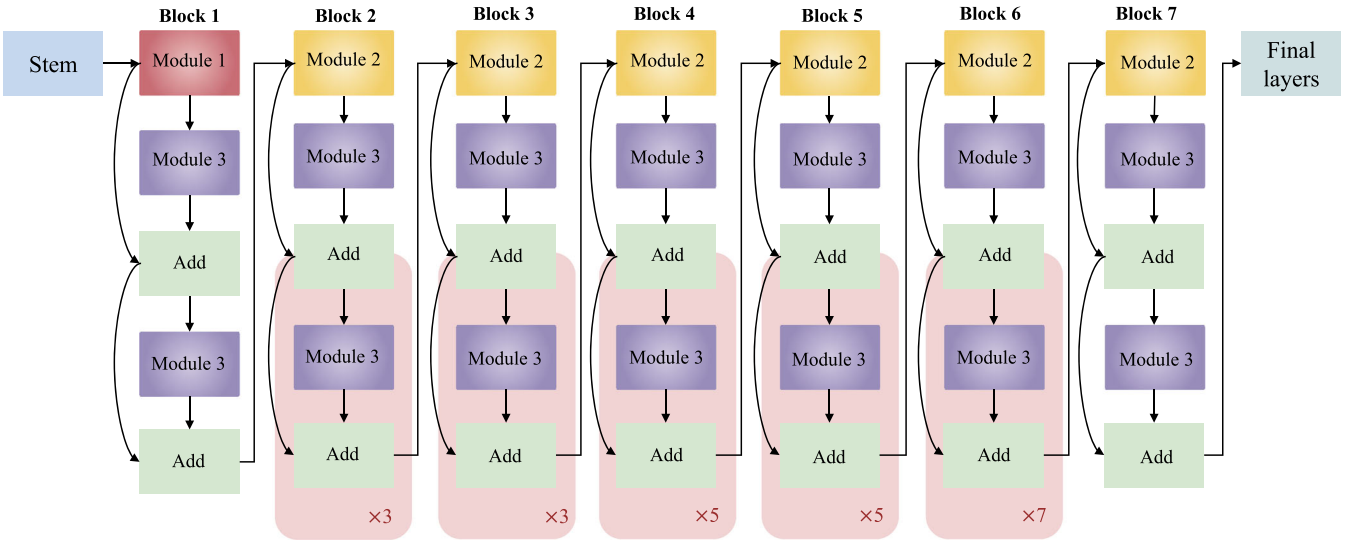


FIGURE 6 Architecture of EfficientNet-B5 [38] (reprinted with permission from V. Agarwal [38], copyright 2020 V. Agarwal).

TABLE 4 Model accuracy on test set.

Set	Number of images	Best case	Worst case	Mean (fivefold)	*SD* (fivefold)
Unseen augmented test set	1008	91.96	90.67	91.29	0.46
Unseen original test set	62	98.39	95.16	96.12	1.43

4 | EXPERIMENTAL RESULTS

In this section, the results obtained from transfer learning are presented and explained in terms of different metrics.

4.1 | Two-tier cross-test and fivefold cross-validation

To strengthen the accuracy and reliability of the model, in addition to the results obtained for the unseen augmented test data on EfficientNet-B5, a fivefold cross-validation was performed on the original test data. The accuracy achieved using this process is listed in Table 4. For comparison, the performance on augmented data is also listed in the table.

For EfficientNet-B5, its weights were used for transfer learning, and fully connected layers were adapted to classify normal and pathological brain MR images. This model achieved high performance on both augmented and real data. The model reached much higher performance over the real MR images than over the augmented MR images. In addition, the model achieved accuracy ranging between 98.39% and 95.16%, with an average accuracy of 96.12% for fewer images, while the accuracy reduced with more images. Two conclusions can be drawn from these results. First, the model performance on a large dataset should be tested with real MR images because rotations alter the images during data augmentation. Second, although the standard deviation of the accuracy in real MR images is only 1.43%, it is a substantial rate for medical imaging. Therefore, the stability of the model must be tested on real images.

4.2 | Confusion matrices

After examining the accuracy, we calculated the confusion matrices on the original and augmented test sets. The confusion matrix provides an in-depth analysis of the accuracy with which each class is predicted. The confusion matrices for both test sets are shown in Figure 7. Each cell is color-coded according to the number of images (number of elements). In other words, as the number of elements in a cell increases, the color of the cells changes from light to dark blue, as indicated by the corresponding color bars. The upper limits of these color bars vary according to the number of images. The confusion matrix provides information on the test set and numbers of correct and incorrect predictions in the classification model. This table contains the true positive, true negative, false positive, and false negative values [35, 36].

The precision, recall, and F1-score can be extracted from the confusion matrix to evaluate the model performance. The precision reflects the positive predictions. The recall is the true positive rate. The F1-score is the harmonic mean of the precision and recall. These metrics were calculated as shown in Table 5 [40].

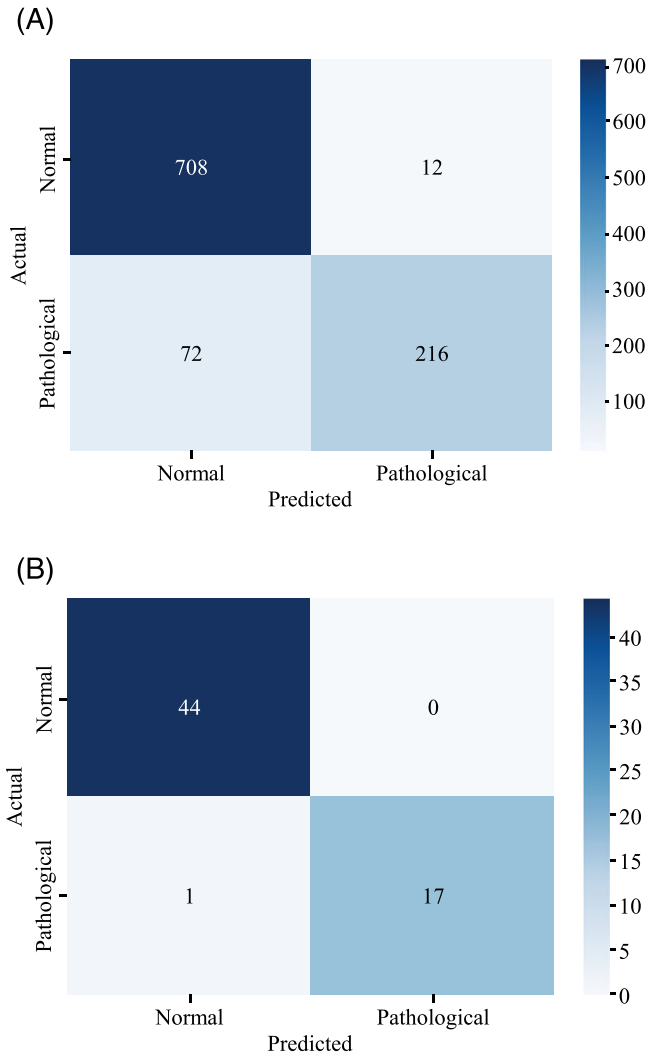


FIGURE 7 Confusion matrices for cross-validation on (A) augmented and (B) original test sets.

As shown in Table 5, the performance on the real dataset was high. The difference in the ratio between precision (100%) and recall (sensitivity) (94.44%) was due to the small sample size. Therefore, additional real data were required. On the augmented dataset, the pathological classes were better classified than the normal classes. The lower performance obtained from the synthetic data confirmed the necessity to use real images.

5 | DISCUSSION

Owing to the use of ML over the past 20 years and success of DL in the past decade, they have been applied to medical image processing. As shown in Table 6, especially over the past 5 years, studies have evolved toward transfer learning using deep architectures. Some studies using pretrained models in Table 6 and the accuracy of our model indicate that pretrained models used to classify different images in the ImageNet competition can achieve high-performance classification of medical images. This finding fulfills RQ1.

Table 6 shows that the proposed transfer learning method based on the pretrained EfficientNet-B5 outperforms other deep models using a DNN [41] or kernel-based ELM [28] in terms of the best-case accuracy for fivefold cross-validation on real MR images. In addition, our method outperforms other models used in transfer learning, such as GoogLeNet [42], VGG19 [44], and ResNet with a randomized neural network [31]. Some models achieved 100% accuracy by using a CNN [45], ResNet34 [43], and AlexNet [29]. However, the comparison models, especially those with deep architectures, may show overfitting. A 100% accuracy rate may indicate that the model memorized the data. Overfitting is a major problem in DL, and it can be mitigated by applying data augmentation. However, data augmentation did not contribute to the classification accuracy in our study, thus answering RQ2. In some DL applications, accuracy is improved when using a large dataset whereas high precision is required in medical data. Consequently, artificial data can adversely affect the classification accuracy. We

TABLE 5 Metrics extracted from confusion matrices.

Metric	Calculation	Original data (%)	Augmented data (%)
Accuracy	(TP + TN)/(TP + TN + FP + FN)	98.39	91.67
Precision	TP/(TP + FP)	100.00	94.74
Recall	TP/(TP + FN)	94.44	75.00
F1-score	2 × (precision × recall)/(precision + recall)	97.14	83.72

Abbreviations: FN, false negative; FP, false positive; TN, true negative; TP, true positive.

TABLE 6 Classification studies on medical imaging.

Study	Method	Accuracy (%)	Number of images
[41]	DNN	97.50	1000 MR images
[28]	Kernel-based ELM	97.04	125 MR images
[29]	Transfer learning with AlexNet	100.00	291 MR images
[42]	Transfer learning with GoogLeNet	98.00	3064 MR images
[43]	Transfer learning with ResNet34	100.00	613 MR images
[44]	Transfer learning with VGG19	94.82	3064 MR images
[31]	ResNet and randomized neural network	95.00	331 MR images
[45]	CNN	100.00	75 MR images

can answer RQ3 as follows. When uncontrolled overfitting and memorization can occur, which is undesirable [46], the network response may differ when an image not contained in the training set is used for testing. To prevent a model from memorizing data and obtain more realistic and reliable results, we conducted a two-tier cross-test, increasing the robustness of the proposed model. This explanation also addresses RQ4, being an important contribution of our study.

Overall, our proposed model stands out for its two-tier cross-test setup, high accuracy on unseen real data, and reliability and high accuracy while avoiding overfitting.

As one limitation of our study, some images did not include all the axial midline structures owing to the operator settings for MRI acquisition. Another limitation was the difficulty in obtaining medical images. Research on real patient data poses difficulties in data availability and obtaining approval from ethics committees.

6 | CONCLUSION AND FUTURE WORK

We propose a transfer learning method for classifying brain MR images as pathological or normal with high accuracy. EfficientNet-B5 achieved an accuracy of 98.39% on real data, indicating that it can be used in clinical studies for decision support regarding brain diseases. Similarly, it achieved an accuracy of 91.67% on synthesized augmented data. Although it is difficult to obtain images of brain diseases, this type of study can help clinicians, especially during examination and diagnosis.

This study is an important step toward creating a general method instead of problem-specific solutions. Segmentation studies are abundant, but different symptoms of brain diseases can lead to different results. Therefore, classification of normal and pathological images is an important first step for reaching a general method. We demonstrated that DL allows to accurately classify brain abnormalities using transfer learning. The predictive power of the model was high, as determined by the classification of unseen test samples that were not included during training. The promising results of the two-tier cross-test showed that the proposed method may provide an alternative and complementary approach for the clinical diagnosis of alterations in the brain structure and function as well as other brain abnormalities. Methods designed to help treat brain diseases can be highly beneficial to healthcare in general, and this study indicates that DL can produce accurate results for medical research.

Transferring deep architectures, which have become more common with the development of AI technologies, to various applications through transfer learning is promising for current and future research. In future work, we will aim to increase the number of pathological images and obtain higher performance on more real images. We will also explore the accuracy of the results obtained by applying different data augmentation operations to real datasets.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ETHICS STATEMENT

The study received the Institutional Ethics Committee approval of the Ankara Training and Research Hospital (date October 27, 2021; approval number 722) was obtained.

INFORMED CONSENT

All the participants provided informed consent before being enrolled in this study.

ORCID

Serkan Savaş https://orcid.org/0000-0003-3440-6271
Çağrı Damar https://orcid.org/0000-0001-7208-1290

REFERENCES

1. Human Brain Facts. Human brain diseases list—causes, symptoms and treatments. 2020. Available from: https://www.humanbrainfacts.org/human-brain-diseases-list.php [last accessed March 2021].
2. C. G. Jennings, R. Landman, Y. Zhou, J. Sharma, J. Hyman, J. A. Movshon, Z. Qiu, A. C. Roberts, A. W. Roe, X. Wang, H. Zhou, L. Wang, F. Zhang, R. Desimone, G. Feng, *Opportunities and challenges in modeling human brain disorders in transgenic primates*, Nat. Neurosci. **19** (2016), 1123–1130. DOI 10.1038/nn.4362.
3. U. R. Acharya, S. L. Fernandes, J. E. WeiKoh, E. J. Ciaccio, M. K. M. Fabell, U. J. Tanik, V. Rajinikanth, and C. H. Yeong, *Automated detection of Alzheimer's disease using brain MRI images—a study with various feature extraction techniques*, J. Med. Syst. **43** (2019), 302. DOI 10.1007/s10916-019-1428-9.
4. N. D. Prins and P. Scheltens, *White matter hyperintensities, cognitive impairment and dementia: an update*, Nat. Rev. Neurol. **11** (2015), 157–165. DOI 10.1038/nrneurol.2015.10.
5. N. Sarbu, R. Y. Shih, R. V. Jones, I. Horkayne-Szakaly, L. Oleaga, and J. G. Smirniotopoulos, *White matter diseases with radiologic–pathologic correlation*, Radiographics **36** (2016), 1426–1447. DOI 10.1148/rg.2016160031.
6. N. S. van der Knaap, R. Schiffmann, F. Mochel, and N. I. Wolf, *Diagnosis, prognosis, and treatment of leukodystrophies*, Lancet Neurol. **18** (2019), 962–972. DOI 10.1016/S1474-4422(19)30143-7.
7. J. M. Wardlaw, M. C. Valdés Hernández, and S. Muñoz-Maniega, *What are white matter hyperintensities made of? Relevance to vascular cognitive impairment*. J. Am. Heart Assoc. **4** (2015), 001140. DOI 10.1161/JAHA.114.001140.
8. S. Debette and H. S. Markus, *The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis*, BMJ **26** (2010), 341. DOI 10.1136/bmj.c3666.
9. H. Y. Hu, Y. N. Ou, X. N. Shen, Y. Qu, Y. H. Ma, Z. T. Wang, Q. Dong, L. Tan, and J. T. Yu, *White matter hyperintensities and risks of cognitive impairment and dementia: a systematic review and meta-analysis of 36 prospective studies*, Neurosci. Biobehav. Rev. **120** (2021), 16–27. DOI 10.1016/j.neubiorev.2020.11.007.
10. A. Khalaf, K. Edelman, D. Tudorascu, C. Andreescu, C. F. Reynolds, and H. Aizenstein, *White matter hyperintensity accumulation during treatment of late-life depression*, Neuropsychopharmacology **40** (2015), 3027–3035. DOI 10.1038/npp.2015.158.
11. M. J. Kempton, J. R. Geddes, U. Ettinger, S. C. Williams, and P. M. Grasby, *Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder*, Arch. Gen. Psychiatry **65** (2008), 1017–1032. DOI 10.1001/archpsyc.65.9.1017.
12. M. Kim and V. Jewells, *Multimodal image analysis for assessing multiple sclerosis and future prospects powered by artificial intelligence*, Semin. Ultrasound CT MRI **41** (2020), 309–318. DOI 10.1053/j.sult.2020.02.005.
13. S. Savaş, N. Topaloğlu, Ö. Kazcı, and P. N. Koşar, *Classification of carotid artery intima media thickness ultrasound images with deep learning*, J. Med. Syst. **43** (2019), 273. DOI 10.1007/s10916-019-1406-2.
14. M. M. Yapıcı, A. Tekerek, and N. Topaloğlu, *Literature review of deep learning research areas*, Gazi Mühendislik Bilimleri Dergisi **5** (2019), 188–215. DOI 10.30855/gmbd.2019.03.01.
15. S. Buyrukoğlu, *Early detection of Alzheimer's disease using data mining: comparison of ensemble feature selection approaches*, Konya J. Eng. Sci. **9** (2021), 50–61. DOI 10.36306/konjes.731624.
16. R. Butuner and M. H. Calp, *Diagnosis and detection of COVID-19 from lung tomography images using deep learning and machine learning methods*, Int. J. Intell. Syst. Appl. Eng. **10** (2022), 190–200.
17. Y. Yılmaz and S. Buyrukoğlu, *Hybrid machine learning model coupled with school closure for forecasting COVID-19 cases in the most affected countries*, Hittite J. Sci. Eng. **8** (2021), 123–131. DOI 10.17350/HJSE19030000222.
18. S. Buyrukoğlu and A. Akbaş, *Machine learning based early prediction of type 2 diabetes: a new hybrid feature selection approach using correlation matrix with heatmap and SFS*, Balkan J. Electr. Comput. Eng. **10** (2022), 110–117. DOI 10.17694/bajece.973129.
19. M. Hanefi Calp, *Use of deep learning approaches in cancer diagnosis*, In *Deep learning for cancer diagnosis. Studies in computational intelligence*, Vol. **908**, Springer, Singapore, 2021, 249–267. DOI 10.1007/978-981-15-6321-8_15.
20. J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, *A proposal for the Dartmouth Summer Research Project on artificial intelligence, August 31, 1955*, AI Mag. **27** (2006), 12. DOI 10.1609/aimag.v27i4.1904.
21. S. Savaş, N. Topaloğlu, Ö. Kazcı, and P. N. Koşar, *Performance comparison of carotid artery intima media thickness classification by deep learning methods*, (International Congress on Human-Computer Interaction, Optimization, and Robotic Applications, Urgup, Turkey), 2019, pp. 125–131. DOI 10.36287/setsci.4.5.025.
22. T. B. Dyrby, E. Rostrup, W. F. Baaré, E. C. van Straaten, F. Barkhof, H. Vrenken, S. Ropele, R. Schmidt, T. Erkinjuntti, L.O. Wahlund, L. Pantoni, D. Inzitari, O.B. Paulson, L.K. Hansen, G. Waldemar, LADIS study group, *Segmentation of age-related white matter changes in a clinical multi-center study*, Neuroimage **41** (2008), 335–345. DOI 10.1016/j.neuroimage.2008.02.024.
23. R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M.C. Valdés-Hernández, D.A. Dickie, J. Wardlaw, D. Rueckert, *White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks*, NeuroImage Clin. **17** (2018), 918–934. DOI 10.1016/j.nicl.2017.12.022.
24. A. Arı and D. Hanbay, *Bölgesel evrişimsel sinir ağları tabanlı MR görüntülerinde tümör tespiti*, Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi **34** (2018), 1395–1408. DOI 10.17341/gazimmfd.460535.
25. P. Roca, A. Attye, L. Colas, A. Tucholka, P. Rubini, S. Cackowski, J. Ding, J.F. Budzik, F. Renard, S. Doyle, E.L. Barbier, I. Bousaid, R. Casey, S. Vukusic, N. Lassau, S. Verclytte, F. Cotton, B. Brochet, R. Casey, F. Cotton, J. de Sèze, P. Douek, F. Guillemin, D. Laplaud, C. Lebrun-Frenay, L. Mansuy, T. Moreau, J. Olaiz, J. Pelletier, C. Rigaud-Bully, B. Stankoff, S. Vukusic, R. Marignier, M. Debouverie, G. Edan, J. Ciron, A. Ruet, N. Collongues, C. Lubetzki, P. Vermersch, P.

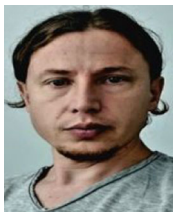
Labauge, G. Defer, M. Cohen, A. Fromont, S. Wiertlewsky, E. Berger, P. Clavelou, B. Audoin, C. Giannesini, O. Gout, E. Thouvenot, O. Heinzlef, A. al-Khedr, B. Bourre, O. Casez, P. Cabre, A. Montcuquet, A. Créange, J.P. Camdessanché, J. Faure, A. Maurousset, I. Patry, K. Hankiewicz, C. Pottier, N. Maubeuge, C. Labeyrie, C. Nifle, R. Ameli, R. Anxionnat, A. Attye, E. Bannier, C. Barillot, D. Ben Salem, M.P. Boncoeur-Martel, F. Bonneville, C. Boutet, J.C. Brisset, F. Cervenanski, B. Claise, O. Commowick, J.M. Constans, P. Dardel, H. Desal, V. Dousset, F. Durand-Dubief, J.C. Ferre, E. Gerardin, T. Glattard, S. Grand, T. Grenier, R. Guillevin, C. Guttmann, A. Krainik, S. Kremer, S. Lion, N. Menjot de Champfleur, L. Mondot, O. Outteryck, N. Pyatigorskaya, J.P. Pruvo, S. Rabaste, J.P. Ranjeva, J.A. Roch, J.C. Sadik, D. Sappey-Marinier, J. Savatovsky, J.Y. Tanguy, A. Tourbah, T. Tourdias, *Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI*, Diagn. Interv. Imaging **101** (2020), 795–802. DOI 10.1016/j.diii.2020.05.009.

26. P. H. B. Diniz, T. L. Valente, J. O. Diniz, A. C. Silva, M. Gattass, N. Ventura, B. C. Muniz, and E. L. Gasparetto, *Detection of white matter lesion regions in MRI using SLIC0 and convolutional neural network*, Comput. Methods Programs Biomed. **167** (2018), 49–63. DOI 10.1016/j.cmpb.2018.04.011.
27. M. Saritha, K. P. Joseph, and A. T. Mathew, *Classification of MRI brain images using combined wavelet entropy based spider web plots and probabilistic neural network*, Pattern Recogn. Lett. **34** (2013), 2151–2156. DOI 10.1016/j.patrec.2013.08.017.
28. S. Lu, Z. Lu, J. Yang, M. Yang, and S. Wang, *A pathological brain detection system based on kernel based ELM*, Multimed. Tools Appl. **77** (2018), 3715–3728. DOI 10.1007/s11042-016-3559-z.
29. S. Lu, Z. Lu, and Y.-D. Zhang, *Pathological brain detection based on AlexNet and transfer learning*, J. Comput. Sci. **30** (2019), 41–47. DOI 10.1016/j.jocs.2018.11.008.
30. A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet classification with deep convolutional neural networks*, Adv. Neural Inf. Process. Syst. **25** (2012), 1097–1105.
31. S. Lu, S.-H. Wang, and Y.-D. Zhang, *Detecting pathological brain via ResNet and randomized neural networks*, Heliyon **6** (2020), e05625. DOI 10.1016/j.heliyon.2020.e05625.
32. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA), 2016, pp. 770–778.
33. Z. U. H. Usmani, *What is Kaggle, why I participate, what is the impact?* 2018, Available from: https://www.kaggle.com/getting-started/44916 [last accessed March 2021].
34. Google Research. *What is Colaboratory?* 2021, Available from: https://research.google.com/colaboratory/faq.html [last accessed March 2021].
35. S. Thrun and L. Pratt, *Learning to learn: introduction and overview*, In *Learning to learn*, Springer, Boston, MA, USA, 1998, 3–17.
36. S. J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Trans. Knowl. Data Eng. **22** (2010), 1345–1359, 2010, 10.1109/TKDE.2009.191.
37. M. Tan and Q. Le, *EfficientNet: rethinking model scaling for convolutional neural networks*, (International Conference on Machine Learning, Long Beach, CA, USA), 2019, pp. 6105–6114.
38. V. Agarwal, *Complete architectural details of all EfficientNet models*, Available from: https://towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142 [last accessed February 2023].
39. S. Savaş, *Detecting the stages of Alzheimer's disease with pre-trained deep learning architectures*, Arab J. Sci. Eng. **47** (2022), 2201–2218. DOI 10.1007/s13369-021-06131-3.
40. F. Uysal, F. Hardalaç, O. Peker, T. Tolunay, and N. Tokgöz, *Classification of shoulder X-ray images with deep learning ensemble models*, Appl. Sci. **11** (2021), 2723. DOI 10.3390/app11062723.
41. A. A. Pandian and R. Balasubramanian, *Performance analysis of texture image retrieval for curvelet, contourlet transform and local ternary pattern using MRI brain tumor image*, Int. J. Found. Comput. Sci. Technol. **5** (2015), 33–46. DOI 10.5121/ijfcst.2015.5604.
42. S. Deepak and P. M. Ameer, *Brain tumor classification using deep CNN features via transfer learning*, Comput. Biol. Med. **111** (2019), 103345. DOI 10.1016/j.compbiomed.2019.103345.
43. M. Talo, U. B. Baloglu, Ö. Yıldırım, and U. R. Acharya, *Application of deep transfer learning for automated brain abnormality classification using MR images*, Cogn. Syst. Res. **54** (2019), 176–188. DOI 10.1016/j.cogsys.2018.12.007.
44. Z. N. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, and J. Lu, *Brain tumor classification for MR images using transfer learning and fine-tuning*, Comput. Med. Imaging Graph. **75** (2019), 34–46. DOI 10.1016/j.compmedimag.2019.05.001.
45. D. R. Nayak, R. Dash, and B. Majhi, *Automated diagnosis of multi-class brain abnormalities using MRI images: a deep convolutional neural network based method*, Pattern Recogn. Lett. **138** (2020), 385–391. DOI 10.1016/j.patrec.2020.04.018.
46. O. Güler and İ. Yücedağ, *Hand gesture recognition from 2D images by using convolutional capsule neural networks*, Arab. J. Sci. Eng. **47** (2022), 1211–1225. DOI 10.1007/s13369-021-05867-2.

AUTHOR BIOGRAPHIES

Serkan Savaş is an associate professor in the Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Kırıkkale University, Turkey. He focused on data mining and conducted research on artificial intelligence, deep learning, cybersecurity, and social networks during his doctoral studies at Gazi University, Turkey. Throughout his career, he has authored and managed numerous national and international projects, including smart maintenance utilizing artificial intelligence, artificial intelligence applications in education, 3D educational technologies, entrepreneurship, and social- and local-purpose projects in the organizations with which he

has been associated. His primary research interests include artificial intelligence, machine learning, deep learning, cybersecurity, data mining, big data, and image processing.



Çağrı Damar, MD, obtained his medical degree from the Faculty of Medicine at Gazi University, Ankara, Turkey, in 1997. He completed his residency in Radiology at Dışkapı Yıldırım Beyazıt Education and Research Hospital, University of Health Sciences, Ankara, Turkey, in 2011. Additionally, between 2013 and 2015, he pursued a fellowship in pediatric radiology at the Faculty of Medicine at Gazi University, Ankara, Turkey. From 2018 to 2020, he served as an assistant professor in the Department of Radiology at Gaziantep University, Gaziantep, Turkey. In December 2021, he joined the Department of Pediatric Radiology at Bikent City Hospital Complex Children's Hospital, University of Health Sciences, Ankara, Turkey, where he currently holds the position of associate professor. His primary research interests include radiologic imaging of humans using various modalities, interventional radiology, and artificial intelligence in radiology.

How to cite this article: S. Savaş and Ç. Damar, *Transfer-learning-based classification of pathological brain magnetic resonance images*, ETRI Journal **46** (2024), 263–276. DOI 10.4218/etrij.2022-0088