Advances in Science and Technology
Research Journal

Advances in Science and Technology Research Journal, 18(9), 204–216
https://doi.org/10.12913/22998624/193627
ISSN 2299-8624, License CC-BY 4.0

Received: 2024.08.24
Accepted: 2024.09.23
Published: 2024.11.01

An Ensemble Transfer Learning Model for Brain Tumors Classification using Convolutional Neural Networks

Bartosz Sterniczuk[1*], Małgorzata Charytanowicz[2]

[1] Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

[2] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

* Corresponding author's e-mail: b.sterniczuk@pollub.pl

ABSTRACT

Convolutional neural networks (CNNs) are a specialized class of deep neural networks. In the present era, these have emerged as highly effective tools for a variety of computer vision tasks. Nonetheless, for classification tasks, the application of a single CNN model is often not sufficient to achieve high precision and robustness. Ensemble learning is a machine learning technique that can improve classification performance through combining multiple models into one. With this method, individual models exchange each other's best performance for each class, resulting in improved overall accuracy. In this work, we studied the performance of CNN models for brain tumor classification. As an outcome, we propose a novel ensemble CNN model for this purpose. We utilized the dataset comes from Nanfang Hospital, which include 3064 MRI images categorized into three types of brain tumor (glioma, meningioma and pituitary). First, we assessed well-known CNN models for their ability to classify brain tumors. Next, we tested several ensemble transfer learning models and created one that utilizes the strengths of the most efficient CNN models. The comparative analysis of model performance demonstrated that the examined ensemble CNN models outperformed all single models. Moreover, regarding evaluation metrics, the proposed model achieved global accuracy of 94% and the highest precision and recall, the F1 score of being 94%. Experimental results revealed that model architecture and ensemble methods have a significant impact on brain tumor classification performance.

Keywords: convolutional neural network, artificial intelligence, brain tumor, ensemble methods, F1-score.

INTRODUCTION

Brain tumor treatment primarily involves surgical resection of the tumor. The surgical treatment to remove the brain tumor depends on the location of the tumor, nature of the lesion and its size [1,2]. Modern surgical techniques allow more precision in such operations, which is crucial for further successful treatment. A factor that increases the chances of survival for patients treated via surgery is early diagnosis of the cancerous process. According to the researchers [2,3,4], by way of better detection of cancer foci, large positive changes can be expected in screening, especially for less common cancers, as well as for patients who have already been treated.

Cancer diagnosis and cancer treatment are two areas that are developing very rapidly. Several experts [5-8] hold the opinion that artificial intelligence is revolutionizing medical research and is playing an increasingly important role in transforming cancer care. The usefulness of artificial intelligence (AI) in oncology is notably due to its ability to process huge amounts of data. Image data has become one of the primary sources of patient information in diagnostics, with a particular focus on cancer diagnosis [9-10]. Medical imaging accounts for almost 90 percent of all healthcare input data, however, deriving a diagnosis from medical image data is very difficult. This is because such data often contain very subtle changes that are difficult to grasp, and hundreds of images need to be reviewed to generate a full diagnosis. The greatest value of computer-aided diagnosis lies in the speed and efficiency of cancer detection.

The second pillar of artificial intelligence algorithm performance is the ability to learn independently [11-13]. The accelerating growth of computational power, the increasing development of machine learning methods, and the broadening of access to diverse medical data have revolutionized AI applications in oncology. Many studies underline the improved accuracy and effectiveness of AI-based systems in interpreting medical images, but more research is needed to understand its full potential and limitations.

Convolutional neural networks (CNNs) are a class of machine learning methods, remarkably effective in image data analysis [14-24]. CNNs can instantly identify cancer foci and assist the doctor in the process of cancer diagnosis. In the hands of a doctor, they can become a powerful assistant, speeding up his/her work and improving its quality. It will also become increasingly better at predicting who might get sick, as well as at more accurately diagnosing the tumor that has already occurred and in foretelling the prognosis of the patients. The approach to patients will be much more individualized as a result.

Neural networks have found applications in the treatment of various conditions beyond image-based diagnosis [25-27]. For instance, in article [28], neural networks were employed to diagnose osteoporosis using vibroarthrography, demonstrating their potential in analyzing non-visual medical data for diagnostic purposes. Another intriguing application is the detection of pulmonary nodules based on local variance analysis and probabilistic neural networks, which highlights the versatility of these systems in identifying and diagnosing complex medical conditions through sophisticated data analysis techniques [29].

The literature offers a wealth of research on ensemble methods, ranging from simpler approaches like soft and hard voting, to more advanced techniques such as weighted majority voting. In the latter, the predicted values are multiplied by pre-determined weights, with the primary focus of research being on how to assign these weights. In our study, we introduce a novel algorithm called F1-WMV, which determines the weights based on the F1-score, offering a new approach to this challenge [30].

This paper is structured as follows. Firstly, related state-of-the-art works and basic concepts are presented in Section 2. Next, Section 3 contains a description of the material and methods that were employed. In Section 4, several well-known CNN models are assessed for recognising brain tumors using MRI image data. Subsequently, the ensemble CNN models were built. Finally, the results obtained, their comparison and conclusions are presented. Section 5 includes a summary of our research and a brief description of future intended research.

RELATED WORKS

The literature review has shown that a number of promising AI-based systems for brain tumor detection are currently being developed. These studies are summarized in Table 1.

The works [31, 34, 35, 36] present the usefulness of data augmentation to improve the performance of CNN models. Sadoon and Ali [32] proposed a novel 28-layer CNN model for brain tumor classification. In their study, the MRI images were pre-processed and augmented to improve the accuracy of the model. The number of slices after augmentation was 15320. The authors achieved very high accuracies of 96.5%, 96.6% and 99.1% for glioma, meningioma and pituitary, respectively, while the overall accuracy of the model was 96.1%. The authors of the article likely achieved better accuracy results through the use of augmentation, a technique not employed in our study. By expanding their dataset through augmentation, they were able to enhance the model's performance, which may explain the observed improvement in accuracy.

In [34], the database from Nanfang Hospital and General Hospital, Tianjin Medical University was investigated using augmentation techniques. The research demonstrated that data expansion techniques significantly boost accuracy, achieving an improvement of up to 96.56%. The final accuracy, after applying augmentation, exceeds the results we achieved. However, the pre-augmentation value is comparable to our outcomes. It is also worth noting that in our study, the training set was divided into three distinct subsets, which may have contributed to weaker model learning.

The study conducted by Saeedi et al. [35] introduced a novel model using the brain tumour dataset, which was compared with an auto-encoder

Table 1. Previous related works and performance comparison in terms of overall accuracy

Author	Database	Accuracy
Wang et al. [31]	BraTS 2018	89.56%
Sadoon and Ali [32]	Nanfang Hospital	96.1%
Dogan and Birant [33]	tested on 28 dataset	90.17%
Badža and Barjaktarović [34]	Nanfang Hospital	96.56%
Saeedi et al. [35]	Nanyang Hospital	96.47%
Sharma et al. [36]	Brain-mri-images	92.0%
Fooladi et al. [37]	MRI scans of glioma	98.64%
Kang et al. [38]	BT-small-2c	94.12%

and six other established machine-learning techniques. Initially comprising 3264 images, the dataset was expanded through augmentation to 9792 samples, with 90% designated for training. The model was tested over 100 epochs with a batch size of 16 and obtained an accuracy of 96.47%.

Sharma and Nandal [36] investigated the combination of the following methods: modified ResNet50, transfer learning and augmentation. Using a modified ResNet50 involved feature extraction and contour cropping of the brain, they achieved 92% accuracy – surpassing competitive frameworks.

In the works [31,37], the authors investigated the impact of data augmentation on brain tumor segmentation. In [31], Wang et al. first obtained Dice scores of 75.70%, 88.98%, and 72.53% for the BraTS 2018 dataset. The use of data augmentation increased the results to 77.70%, 89.56% and 73.04%, respectively. The work [37] proposed a new method of segmentation of brain tumours. They combined the CNN and fuzzy *K*-means algorithms. Their algorithm was tested on the BRATS database and achieved an accuracy of 98.64%.

Ensemble techniques were investigated in [33,38]. Dogan and Birant [33] proposed a novel approach to ensemble learning called Weighted Majority Voting Ensemble, and introduced a novel algorithm for determining weights. In the evaluation of this method, experiments were conducted across 28 widely recognized datasets. The findings indicated that this approach outperformed existing methods in 27 out of 28 cases, demonstrating significant improvements in ensemble learning efficacy. The benefits of the ensemble method were also illustrated by Kang et al. [38]. The authors assessed a set of pre-trained networks and chose the top three based on their performance. These top-performing networks were combined into a single ensemble of deep features, which were then used as input for various classifiers to produce the final prediction. The results of their experiments showed that the proposed method brought the expected results.

In the preceding paragraphs, a variety of experiments on brain tumors and ensemble methods were discussed. The authors aimed to enhance the Weighted Majority Voting method by introducing an innovative approach to weight determination. This new approach distinguishes itself from existing methods by utilizing the F1-Score, which had not been previously applied for this purpose. The proposed solution has the potential to become a valuable tool for researchers, applicable to other datasets and explorations with different machine learning techniques.

MATERIAL AND METHODS

Study material

The MRI image data was collected from 233 patients at the Nanfang Hospital and General Hospital, Tianjin Medical University between 2005 and 2010. The dataset contained 3064 axial, coronal and sagittal plane images with T1-weighted contrast, and included three types of brain tumor: glioma, meningioma and pituitary. Their numbers were 1426, 708 and 930 slices, respectively. The images, with a resolution of 512×512, were originally saved in the MATLAB data format (.mat file). Each row, labeled with the patient ID, contained the following fields describing a tumor [55]:

- cjdata.label – tumor label: 1 – meningioma, 2 – glioma, 3 – pituitary,
- cjdata.PID – patient ID,
- cjdata.image – a matrix (array) containing image data,
- cjdata.tumorBorder – a vector of point coordinates on tumor border,
- cjdata.tumorMask – a binary image representing tumor region.

The procedure of creating the image dataset from a MATLAB file was performed in Python [39] and consisted of the following steps:

```
Algorithm 1: Creating the image dataset
    Input: the MATLAB (.mat) file
    Output: the labelled image dataset
1. For each row of the MATLAB file:
    − Extract a tumor label,
    − Extract an array containing image data,
    − Convert array values into the range between 0 and 255,
    − Create an image from the array using the function Image.fromarray(array),
2. Return the labelled image dataset.
```

Figure 1 shows exemplary axial, coronal and sagittal slices of three types of brain tumor [39].

Model evaluation

The performance of classification models can be assessed by utilizing confusion matrices. Here, the size of the confusion matrix is equal to $c \times c$, where c is the number of classes, and the diagonal of the matrix displays the numbers n_{ij}, $i = 1, 2, \ldots, c$, of correctly classified elements for class i. The remaining fields contain the numbers of misclassified elements in relation to their actual and predicted labels. The sum of all values n_{ij}, $j = 1, 2, \ldots, c$, is equal to the number of all elements N [40]. The confusion matrix is interpreted using the following terms:

- True Positive (TP) for a class – a diagonal value of the corresponding row and column,
- False Positive (FP) for a class – the sum of values of the corresponding column except for that of the diagonal value,

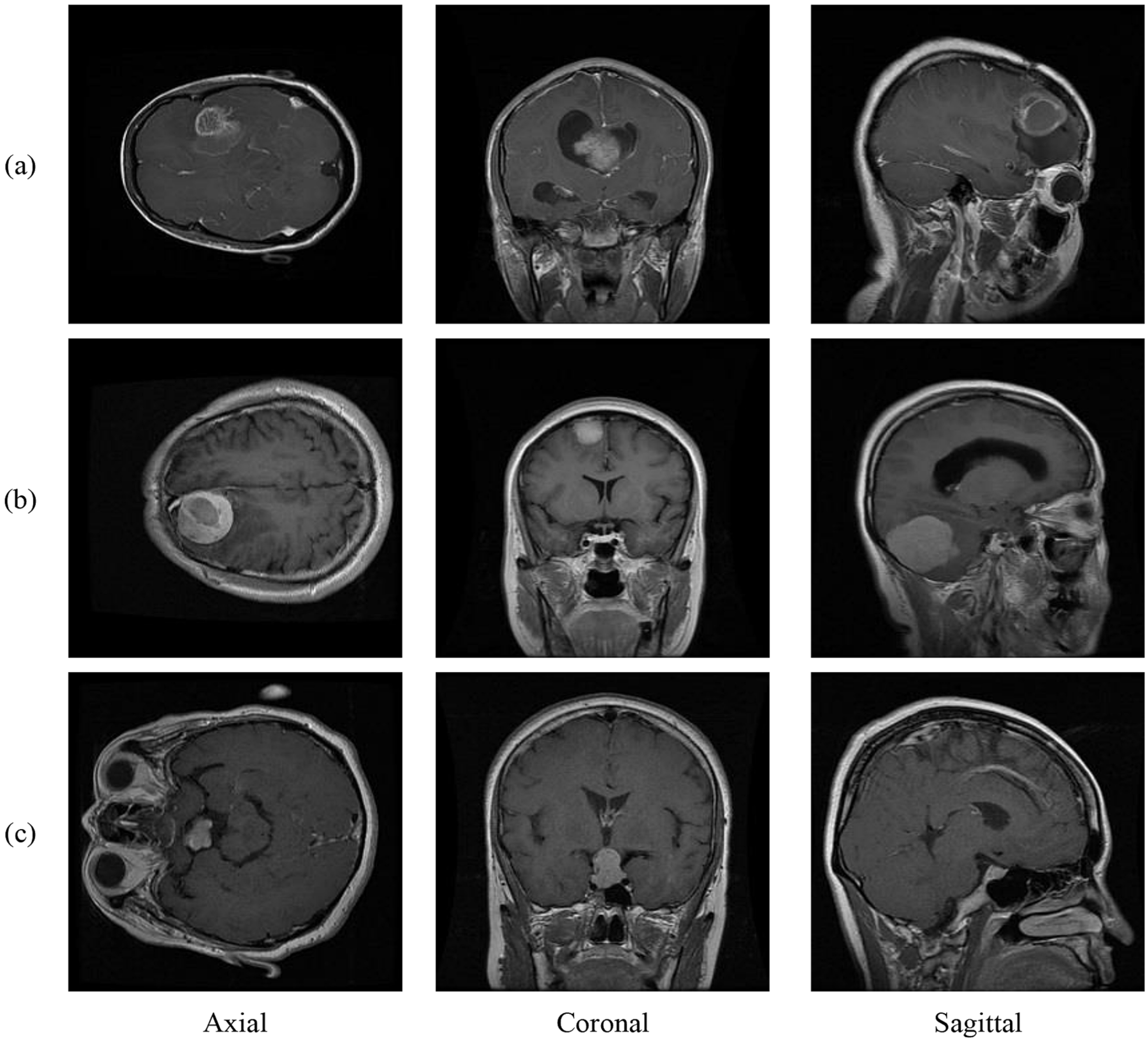

Figure 1. Axial, coronal and sagittal MRI slices of three types of brain tumor: (a) glioma, (b) meningioma, (c) pituitary

- False Negative (FN) for a class – the sum of values of the corresponding row except for that of the diagonal value.
- True Negative (TN) for a class – the sum of values of all columns and rows except for that of the values of the corresponding row and column.

A three-class confusion matrix is given in Table 2. Thus, $TP_B = n_{22}$, $FP_B = n_{12} + \mathrm{n}_{32}$, $FN_B = n_{21} + \mathrm{n}_{23}$, $TN_B = n_{11} + n_{13} + n_{31} + n_{33}$.

The base evaluation metrics for class *k* are defined using the confusion matrix and the following evaluation metrics: accuracy, precision, recall and F1-score. They are defined as follows:

Accuracy is the rate of correctly classified elements to the class among all elements:

$$Accuracy_k = \frac{TP_k+TN_k}{TP_k+FN_k+FP_k+TN_k}, \quad (1)$$

Precision is the rate of correctly classified elements to the class *k* among all elements assigned to this class by a model:

$$Precision_k = \frac{TP_k}{TP_k+FP_k}, \quad (2)$$

Recall is the rate of correctly classified elements to the class *k* among all elements of this class:

$$Recall_k = \frac{TP_k}{TP_k+FN_k}, \quad (3)$$

F1 score is a harmonic mean of recall and precision:

$$F1score_k = \frac{2}{\frac{1}{Precision_k}+\frac{1}{Recall_k}}. \quad (4)$$

Evaluation metrics computed for each class can be combined using macro- or micro averaging to describe overall performance of a model [13, 39].

The classification models were evaluated using *k*-fold cross-validation, ensuring the stability of the classification results. In this approach, the data is divided into *k* subsets, where each subset is used once as a validation set, while the remaining *k*-1 subsets are used as a training set. This process is repeated *k* times, and the evaluation metrics are averaged [41, 42]. In our study, we tested *k* value of 10, which are commonly used in similar research. It is important to note that a smaller *k* results in higher variance of the classification error, while a larger *k* increases the model's bias. A compromise is usually needed, with *k* values of 5 or 10 being most common. For *k* = 10, the classification error rate is stable [43].

CNN models

Convolutional Neural Networks (CNNs) are an advanced type of artificial intelligence. They are used especially in spatial processing (i.e. images and videos). By way of their construction, they enable automatic and adaptive learn hierarchical data representations. Hence, employment of CNNs enables extremely effective image recognition. Typically, CNNs use three types of layers: convolutional, pooling and fully connected. The main goal of the convolutional layer is to extract relevant information from the image, such as edges, textures and more sophisticated patterns. The key to this algorithm effectiveness is the use of a filter (sometimes called a 'kernel'), which moves over data, generating a splot operation. Another important process occurring during the learning phase of CNN exploitation is the dimension reduction of feature maps, which is implemented in the pooling layer. The method chooses the maximum value from the feature maps, reducing the data and increasing the robustness of the model. The last layer – fully connected (also called 'dense'), is responsible for the final classification. Each neuron in each layer is connected to every neuron in the previous layer, allowing global integration of the extracted features [44].

VGG16 stands out as one of the most widely employed pre-trained convolutional neural networks for image classification. According to its creators, the model can classify 1000 images across various categories with an impressive accuracy of 92.7%. This high level of accuracy is largely attributed to the use of small convolutional filters (3×3) with a stride of 1. Although such parameters demand significant computational power, this challenge is mitigated because the model has already undergone pre-training. The

Table 2. A three-class confusion matrix with distinguished class B

Class	Predicted class *A*	Predicted class *B*	Predicted class *C*
Actual class *A*	n_{11} (TN_B)	n_{12} (FP_B)	n_{13} (TN_B)
Actual class *B*	n_{21} (FN_B)	n_{22} (TP_B)	n_{23} (FN_B)
Actual class *C*	n_{31} (TN_B)	n_{32} (FP_B)	n_{33} (TN_B)

name VGG16 reflects the 16 convolutional layers that form the core of its architecture [45]. In the literature, VGG16 is confirmed to be effective for tumour diagnosis. The authors of the study [46] reported about 95% accuracy in detecting brain tumours via MRI scans. This result could lay the groundwork for further research into ensemble methods, potentially enhancing diagnostic accuracy and model robustness.

The VGG19 network closely resembles VGG16, with the primary distinction being the number of convolutional layers. VGG19 includes 19 convolutional layers, three more than its predecessor. This increase in layers leads to a higher number of parameters to be trained, thereby enhancing the network's capability to capture and represent more complex features [45]. In the paper [47], the authors achieved a very high accuracy rate of 99.5% with the VGG19 model, based on brain tumour assessment. This impressive result holds promising potential for further development of this research into ensemble methods.

InceptionV3 is a pre-trained convolutional neural network created by engineers at Google. This model secured victory in the 2015 "ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)" for achieving the highest accuracy in image classification. The remarkable accuracy of InceptionV3 is due to its use of multiple convolutional layers, each employing different kernel sizes. These kernels are notably small, with dimensions such as (1×1, 3×3, 5×5). The authors opted to use the third version of Inception due to its improved capacity compared to earlier iterations. Despite its enhanced capabilities, InceptionV3 maintains a similar speed to the previous versions, making it an efficient choice for complex tasks [48].

ResNet152V2 and ResNet50V2, named after their "Residual Network" architecture, were introduced in 2016 as advanced versions of the original ResNet model. The number 152 in ResNet152V2 signifies that the model comprises 152 layers, while ResNet50V2 is characterized by its 50 layers. Both networks achieve outstanding performance by employing compact filter sizes of 1×1, 3×3, and 1×1.

InceptionResNetV2, developed by Google researchers in 2016, represents an extension of the established Inception and InceptionV3 models. This network integrates multiple filter sizes (1×1, 3×3, 5×5) to optimize both model complexity and computational efficiency, aiming for superior performance outcomes. The article [49] indicates that the Inception ResNetV2 network achieved an accuracy of 93.4%, making it a strong candidate for applying ensemble methods to further improve performance.

DenseNet201, developed by researchers at Facebook, was introduced for commercial use in 2017. It aims to enhance the earlier DenseNet model. The designation "201" in its name indicates the number of layers implemented in the model. This network family is designed specifically for image object recognition. The article [50] achieved an accuracy of 88% during tests on the test dataset. Therefore, it is worth investigating whether the use of ensemble methods could help achieve better results when combined with other models.

The Xception network was proposed by François Chollet in 2017. It utilizes depth-wise separable convolutions, a technique introduced by the author. Compared to Inception, Xception has a simpler structure, which facilitates its implementation and improves its efficiency. The authors chose to use this network in their research based on scientific studies showing that Xception outperforms Inception on the ImageNet dataset [51, 52].

MobileNetV2 is a neural network developed by Google and introduced in 2018. The network consists of approximately 3 million parameters and features around 53 layers. Key innovations include the use of depth-wise separable convolutions, which reduce the number of required computations and enhance performance. In the article [52], a comparison of three networks for detecting brain tumors is presented, with MobileNet emerging as the most effective. Consequently, the authors of the article have decided to include the MobileNet model in their exploration of ensemble methods.

Ensemble models

Ensemble learning is a machine learning technique that improves classification performance by combining multiple models. With this method, individual models exchange each other's best results for each class, resulting in greater overall accuracy compared to a single model [53]. The main aspects of ensemble methods are the following:

- improved stability and accuracy – employing ensemble methods reduces the standard variation and enhances effectiveness of use,
- combining patterns – once predictions have been acquired from each model, the results are

combined in a determined way according to the chosen ensemble technique,

- diversity of models – this method employs different types of models or dissimilar hyperparameters during training. Due to this, the models might commit different mistakes, but by combining the models, the errors can be reduced.

In academic literature, various ensemble methods are extensively discussed. Among these, the more widely recognized techniques include bagging, boosting, stacking, voting and blending. For our study, we specifically employed the voting method exclusively.

Formally, we assume that we have m base models: $f_1, f_2, f_3, \ldots, f_m$. The final prediction $\hat{y}$ might be expressed using the rule (5):

$$\hat{y} = \varphi\big(f_1(x), f_2(x), \ldots, f_m(x)\big) \tag{5}$$

where: the φ is an aggregation function, and x is the input data [54].

'Soft voting' is regarded as one of the most straightforward and widely adopted ensemble methods. This technique involves combining predictions from multiple classifiers by averaging their outputs. The aggregation function can be succinctly expressed via equation (6):

$$\hat{y} = \frac{1}{m}\sum_{i=i}^{m} f_i(x) \tag{6}$$

Another popular approach is 'majority voting' (also known as 'hard voting'), which involves aggregating predictions from multiple base models by selecting the class c that receives the highest number of votes among them. Formally, the decision-making process can be summarized as:

$$\hat{y} = argmax_c \sum_{i=1}^{m}(f_1(x) = c) \tag{7}$$

'Weighted sum' is a method where the probability returned by models is multiplied by their weights. There are a number of ways of determining the coefficients in applying this method, among others, RRMSE, Bagging and AdaBoost [41, 42]. The general formula is:

$$\hat{y} = \sum_{i=1}^{m} \alpha_i f_i(x), \sum_{i=1}^{m} \alpha_i = 1 \tag{8}$$

where: α_i is the coefficient assigned to f_i model.

In literature, very sophisticated ways to combine outputs can be found. One is an adaptive weighted voting fusion recognition algorithm based on entropy [41]. This algorithm depends on calculating the entropy of each sample x. Here, we assume that $p_{i,j}$ is the posterior probability output for i classifier and j class, Hence, we can calculate the entropy with the rule:

$$\alpha_i(x) = -\sum_{j=1}^{c} p_{ij} log_2 p_{i,j}, i = 1,2, \ldots, m \tag{9}$$

Herein, the p_i value is the measure of uncertainty of classifier i. In our paper, we also propose applying fusion weight techniques. Thus, the previously determined coefficient are process by formula (10):

$$\alpha_i = \frac{exp\,(-w_i(x))}{\sum_{j=1}^{m} exp\,\big(-w_j(x)\big)} \tag{10}$$

To determine the final predictions for sample, we need to multiply weights by appropriate probability according to the formula (11):

$$P(x) = \begin{cases} \alpha_1 p_{1,1}(x) & \alpha_1 p_{1,2}(x) & \cdots \alpha_1 p_{1,c}(x) \\ \alpha_2 p_{2,1}(x) & \alpha_2 p_{2,1}(x) & \cdots \alpha_2 p_{2,c}(x) \\ \vdots & \vdots & \vdots \\ \alpha_m p_{m1}(x) & \alpha_m p_{m2}(x) & \cdots \alpha_m p_{m3}(x) \end{cases} \tag{11}$$

Finally, we are able to determine the class affiliation. The process depends on selecting the classes with the highest probability. The formal course of action follows rule (12):

$$\hat{y} = argmax_c \sum_{i=1}^{c} \alpha_i p_{ij} \tag{12}$$

A novel three-steps approach, called a 'Weighted Majority Voting Ensemble' (WMVE) that relies on incorrect classifications was presented in [33]. In applying this approach, in the initial phase, the classifiers are trained on the test dataset. In the second stage, classifiers undergo training using the validation set, with initial weights set uniformly to one. The weights of classifiers that accurately predict the class label of an instance are then adjusted based on the ratio of classifiers making incorrect predictions to the total number of classifiers. For example, for three classifiers $f_1(x), f_2(x), f_3(x), y\{0,1\}$ and their outputs 0, 0, 1, the actual class is 1. The weights are $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 1 + 2/3 = 1.67$. Next, the final result is obtained by applying equation (8).

Proposed CNN ensemble model

The proposed classifier combines a few classifiers based on the weighted majority voting approach. Weight parameters were added to base classifiers according to their performance given by F1-score metrics. For this purpose, F1-scores of each class computed on the validation set were assigned to each classifier. Next, these scores were multiplied by the corresponding predicted class-probabilities of each base classifier. The weighted average was subsequently assigned to each class and the class with highest score result was selected. The whole procedure can be written as follows:

```
Algorithm 2. F1-weighted majority voting
Input: the labelled image dataset $X$
Output: the final predicted class for each testing sample $x$ in the test set
Initial values and control parameters:
    $m$ – number of machine learning models,
    $c$ – number of brain tumor classes,
    $k$ – number of folds.
# Phase 1. Compute F1-scores
1. Choose $m$ base machine learning classifiers $f_1, f_2, f_3, \dots, f_m$.
2. Read the labelled image dataset $X$ containing data of $c$ classes of brain tumor.
3. Split $X$ into train, validation and test sets.
4. For each model $f_i$, $i = 1, \dots, m$:
    4.1. Compute $F1$-score denoted by $F_{ij}$ for each class $j$, $j = 1, \dots, c$ on the validation set using $k$-fold cross-validation.
# Phase 2. Predict the labels in the test set
5. For each testing sample $x$ in the test set:
    5.1. For each model $f_i$, $i = 1, \dots, m$:
         Compute class probabilities  for  on the element $x$.
    5.2. For each class $j$, $j = 1, \dots, c$:
         Compute the F1-weighted average predicted value $pv_j = \sum_{i=1}^{m} F_{ij} p_{ij}$.
    5.3. Determine the class with the largest $pv_j$, $j$, $j = 1, \dots, c$ for the testing sample $x$ using the rule $finalpredictedclass(x) = argmax\ (pv_j)$.
6. Return the final predicted classes for testing samples in the test set.
```

Computing environment

All research was done using the Python programming language with the aid of Jupiter notebooks. The following libraries were exploited:

- keras – for preparing previously pre-trained networks,
- matplotlib – to generate plots,
- scikit-learn – with the aim of using evaluating metrics,
- tenserflow – to enable the use of GPUs.

The training was carried out on a computer with the following specifications:

- Processor – Intel Core i5-9400F,
- GPU – RTX 2060Ti,
- RAM – 16 GB DRR4.

EXPERIMENTS AND RESULTS

In our research, we investigated several CNN classification models, as well as ensemble models. Table 3 and Table 4 contained the CNN model characteristics and hyperparameter values used in our experiments.

The brain tumor dataset BraTS including 3064 axial, coronal and sagittal plane images, was divided as follows. The test set contained 20% of the total number of samples and was randomly selected. The remaining 80% was used to create training and validation sets (Table 5). In our experiments, we applied k-fold cross-validation with $k = 10$.

Table 6 and Table 7 show a comparison of the performance of examined CNN classification models using 10-fold cross-validation. The best results obtained for the experiment are bolded.

We found that the accuracy ranged from 86.84% to 94.99% and from 85.46% to 92.73% for the training and test sets, respectively. This confirms the very high effectiveness of the CNN models when applied to brain tumor detection. The highest accuracy was achieved by applying the ResNet50V2 model, the lowest – by the VGG19 model.

Table 3. CNN models characteristics

Model	Number of parameters	Number of trainable parameters
VGG16	27,561,795	12,847,107
VGG19	32,871,491	12,847,107
Xception	87,972,395	67,110,915
InceptionV3	59,553,571	37,750,787
ResNet50V2	90,675,715	67,110,915
DenseNet201	66,492,995	48,171,011
InceptionResNetV2	82,650,339	28,313,603
ResNet152V2	125,442,563	67,110,915
MobileNetV2	44,203,075	41,945,091
DenseNet121	32,729,667	25,692,163

Table 4. Hyperparameter values

Parameter	Value
Image size	250 x 250 x 3
Optimization method	RMSprop
Learning rate	$2 \cdot 10^{-5}$
Number of epochs	30
Batch size	16
Dense Layer activation function	relu
Output Layer activation function	softmax
Loss function	categorical_crossentropy

Table 5. Splitting dataset – training and test sets

Brain tumor type	Training (10-fold cross-validation)	Test	Total
Glioma	1140	286	1426
Meningioma	565	143	708
Pituitary	744	186	930
Total	2449	615	3064

Table 6. Comparison of CNN models: experimental results of application upon the training dataset (10-fold cross-validation)

Model	Accuracy		Loss	
	Mean	Std	Mean	Std
VGG16	0.8934	0.0026	0.2651	0.0051
VGG19	0.8684	0.0048	0.3213	0.0066
Xception	0.9405	0.0027	0.1604	0.0078
InceptionV3	0.9168	0.0056	0.2155	0.0137
ResNet50V2	**0.9499**	**0.0058**	**0.1518**	**0.0164**
DenseNet201	**0.9487**	**0.0033**	**0.1377**	**0.0063**
InceptionResNetV2	0.9167	0.0040	0.2233	0.0102
ResNet152V2	0.9441	0.0040	0.1701	0.0123
MobileNetV2	0.9441	0.0027	0.1588	0.0139
DenseNet121	0.9385	0.0055	0.1623	0.0101

Table 7. Comparison of CNN models: experimental results of application upon the validating dataset (10-fold cross-validation)

Model	Accuracy		Loss	
	Mean	Std	Mean	Std
VGG16	0.8828	0.0196	0.3036	0.0366
VGG19	0.8546	0.0303	0.3611	0.0580
Xception	0.9089	0.0151	0.2814	0.0525
InceptionV3	0.8910	0.0213	0.3189	0.1052
ResNet50V2	**0.9273**	**0.0106**	**0.2644**	**0.0616**
DenseNet201	**0.9147**	**0.0196**	**0.2357**	**0.0518**
InceptionResNetV2	0.8881	0.0207	0.3213	0.0706
ResNet152V2	0.9016	0.0212	0.3845	0.0741
MobileNetV2	0.8987	0.0156	0.2945	0.0102
DenseNet121	0.9110	0.0193	0.2369	0.0521

Table 8 presents the basic evaluation metrics: precision, recall and F1-score. All examined CNN models achieved high outcomes, exceeding 80%. Summarizing, the highest scores were achieved by five models: ResNet50V2, DenseNet201, ResNet152V2, MobileNetV2 and DenseNet121.

We assumed that the CNN models used to build the ensemble model can be selected with respect to the best scores for each of the brain tumor types. Table 9 presents the confusion matrices of the examined CNN models. Abbreviations G (glioma), M (meningioma), P (pituitary) are used for clarity, being the short form of the corresponding labels. Accordingly, the highest result of 97% was achieved in pituitary classification and the lowest result of 83% in meningioma classification.

Finally, the examined ensemble models combined three classifiers: MobileNetV2, ResNet50V2 and ResNet152V2. The following ensemble methods were built for this purpose: soft voting, hard voting, entropy, WMVE [33] and the proposed Algorithm 2. Table 10 show basic evaluation metrics obtained by the examined ensemble CNN models. The models were trained

Table 8. Basic evaluation metrics obtained by application of the examined CNN models upon the validating set (10-fold cross-validation)

Model	Precision	Recall	F1 score
VGG16	0.8776	0.8738	0.8724
VGG19	0.8592	0.8510	0.8425
Xception	0.9016	0.9020	0.9120
InceptionV3	0.9017	0.8951	0.8953
ResNet50V2	**0.9203**	**0.9187**	**0.9182**
DenseNet201	**0.9211**	**0.9179**	**0.9169**
InceptionResNetV2	0.8987	0.8930	0.8912
ResNet152V2	**0.9088**	**0.9061**	**0.9055**
MobileNetV2	**0.9141**	**0.9122**	**0.9110**
DenseNet121	**0.9119**	**0.9026**	**0.9072**

Table 9. Confusion matrices of the examined CNN models (10-fold cross-validation), expressed as percentages

VGG16

	G	M	P
G	88.89	10.09	2.02
M	17.01	74.85	8.14
P	1.21	2.70	96.10

VGG19

	G	M	P
G	89.91	8.16	1.93
M	2.82	63.27	8.49
P	3.37	2.29	94.34

Xception

	G	M	P
G	82.37	6.75	0.8
M	10.09	74.26	5.65
P	1.34	4.58	84.08

InceptionV3

	G	M	P
G	89.91	9.12	0.96
M	13.25	81.10	5.65
P	1.35	3.36	95.29

ResNet50V2

	G	M	P
G	93.86	5.70	0.4
M	11.50	**82.67**	5.84
P	0.94	3.23	95.83

DenseNet201

	G	M	P
G	93.86	5.0	1.14
M	12.73	81.80	5.47
P	1.07	2.68	96.24

InceptionResNetV2

	G	M	P
G	90.70	8.25	1.05
M	15.08	78.03	6.89
P	1.62	2.70	95.68

ResNet152V2

	G	M	P
G	91.32	7.87	0.88
M	12.36	81.11	6.53
P	0.81	2.42	**96.77**

MobileNetV2

	G	M	P
G	**95.26**	4.56	0.18
M	17.02	77.16	5.47
P	1.21	3.09	96.24

DenseNet201

	G	M	P
G	93.86	5.0	1.14
M	12.73	81.80	5.47
P	1.07	2.68	9.24

Table 10. Basic evaluation metrics obtained by the examined ensemble CNN models on the test set

Ensemble model	Accuracy	Precision	Recall	F1 score
Soft voting	0.9414	0.9326	0.9313	0.9317
Hard voting	0.9349	0.9260	0.9220	0.9234
Entropy	0.9381	0.9378	0.9381	0.9369
WMVE	0.9397	0.9391	0.9397	0.9387
F1-WMVE	**0.9414**	**0.9411**	**0.9414**	**0.9402**

on the training set. The results are presented as applied upon the independent test set.

We found that there was a significant improvement for all evaluation metrics when the ensemble approach was applied. The worst results, above 93% for all metrics, were obtained by utilizing the simple hard voting classifier. Soft voting and our proposed method gave the same accuracy of 94.14%, but the proposed method gave the best results – exceeding 94% for precision, recall and F1-score.

CONCLUSIONS

The objective of our work has been successfully achieved, and our experiment demonstrates that the F1-WMVE algorithm yielded satisfactory results, outperforming classical methods such as soft and hard voting, entropy and WMVE. It is worth noting that the literature reviewed achieved higher results on the same dataset through augmentation. The proposed method may serve as a foundation for further research, such as expanding the dataset using augmentation techniques.

The results can be summarized as follows:

1) All examined CNN models achieved high evaluation metrics. The highest scores (exceeding 90%) were achieved utilizing ResNet50V2, DenseNet201, ResNet152V2, MobileNetV2 and DenseNet121.
2) The application of the ensemble approach resulted in an increase of the evaluation metrics by an average of 2%.
3) The highest results were achieved by employing the proposed ensemble model. All evaluation metrics exceeded 94%.

In our experiments, the ensemble model combined the three best models as discerned in evaluating each class of brain tumor. Further experiments will be conducted on the selection of base models and the application of F1-score.

Acknowledgements

We would like to extend our heartfelt thanks to Nanfang Hospital and General Hospital, Tianjin Medical University for the opportunity to use the database that was utilized during our research.

REFERENCES

1. Ene C.I., Ferguson S.D. Surgical Management of Brain Metastasis: Challenges and Nuances. Front Oncol 2022 Mar 14; 12. https://doi.org/10.3389/fonc.2022.847110
2. Crosby D., Bhatia S., Brindle K.M., Coussens L.M., Dive C., Emberton M., Esener S., Fitzgerald R.C., Gambhir S.S., Kuhn P., Rebbeck T.R., Balasubramanian S. Early detection of cancer. Science 2022 Mar 18; 375(6586). https://doi.org/10.1126/science.aay9040
3. Beane J., Campbell J.D., Lel J., Vick J., Spira A. Genomic approaches to accelerate cancer interception. Lancet Oncol 2017 Aug; 18(8): e494-e502. https://doi.org/10.1016/S1470-2045(17)30373-X
4. Hossain T., Shishir F.S., Ashraf M., Al Nasim M.A., Muhammad Shah F. Brain Tumor Detection Using Convolutional Neural Network. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Bangladesh, Dhaka, May 3-5 2019. IEEE; 2019. https://doi.org/10.1109/icasert.2019.8934561
5. R. Tamilaruvi, R. Vijayalakshmi, M. Ganthimathi, R. Surendiran, M. Thangamani, S. Satheesh. Brain Tumor Detection in MRI Images using Convolutional Neural Network Technique. SSRG International Journal of Electrical and Electronics Engineering 2022 Dec 31; 9(12): 198-208. https://doi.org/10.14445/23488379/IJEEE-V9I12P118
6. Vollmuth P., Foltyn M., Huang R.Y., Galldiks N., Petersen J., Isensee F., et al. AI-based decision support improves reproducibility of tumor response assessment in neuro-oncology: an international multireader study. Neuro-Oncology 2022 Aug 2; 25(3) : 533-545. https://doi.org/10.1093/neuonc/noac189
7. Sollini M., Bartoli F., Marciano A., Zanca R., Slart R.H., Erba P.A. Artificial intelligence and hybrid imaging: the best match for personalized medicine in oncology. Eur J Hybrid Imaging 2020; 4(1). doi: 10.1186/s41824-020-00094-8.
8. Khalighi S., Reddy K., Midya A., Pandav K.B., Madabhushi A., Abedalthagafi M. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. NPJ Precis Oncol 2024 Mar 29; 8(1). https://doi.org/10.1038/s41698-024-00575-0
9. Yadav S.S., Jadhav S.M. Deep convolutional neural network based medical image classification for disease diagnosis. J Big Data 2019 Dec; 6(1). https://doi.org/10.1186/s40537-019-0276-2
10. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv 2014; 1409.1556.
11. O'Sullivan B., Woodridge M. Artificial Intelligence: Foundations, theory, and algorithms. Book Series, 2015-2021; (9).
12. Aggarwal C.C. Neural Networks and Deep Learning. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-94463-0
13. Chollet, F. Deep Learning with Python. Chollet & Dysart - Manning Publications, 2021.
14. Archana R., Jeevaraj P.S. Deep learning models for digital image processing: a review. Artif Intell Rev 2024; 57(1). https://doi.org/10.1007/s10462-023-10631-z
15. Abbas S., Alhwaiti Y., Fatima A., A Khan M., Adnan Khan M., M Ghazal T., Kanwal A., Ahmad M., Sabri Elmitwally N. Convolutional Neural Network Based Intelligent Handwritten Document Recognition. Comput Mater Amp Contin 2022; 70(3): 4563-81. https://doi.org/10.32604/cmc.2022.021102
16. Litjens G., Kooi T., Bejnordi B.E., Setio A.A., Ciompi F., Ghafoorian M., van der Laak J.A., van Ginneken B., Sánchez C.I. A survey on deep learning

in medical image analysis. Medical image analysis 2017 Dec; 42: 60-88. https://doi.org/10.1016/j.media.2017.07.005

17. Almabdy S., Elrefaei L. Deep Convolutional Neural Network-Based Approaches for Face Recognition. Appl Sci 2019 Oct 17; 9(20): 4397. https://doi.org/10.3390/app9204397
18. Charytanowicz M., Kowalski P.A., Lukasik S., Kulczycki P., Czachor H. Deep learning for porous media classification based on micro-ct images. In: 2022 international joint conference on neural networks (IJCNN), Padua, Italy Jul 18-23 2022. IEEE. https://doi.org/10.1109/ijcnn55064.2022.9891899
19. Łukasik E., Charytanowicz M., Miłosz M., Tokovarov M., Kaczorowska M., Czerwiński D., Zientarski T. Recognition of handwritten Latin characters with diacritics using CNN. Bulletin of the Polish Academy of Sciences: Technical Sciences 2021; 69: 1–12.
20. Fabijańska A., Danek M., Barniak J. Wood species automatic identification from wood core images with a residual convolutional neural network. Comput Electron Agric 2021 Feb; 181:105941. https://doi.org/10.1016/j.compag.2020.105941
21. He K., Zhang X., Ren, S., Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.
22. Su C., Xu S.J., Zhu K.Y., Zhang X.C. Rock classification in petrographic thin section images based on concatenated convolutional neural networks. Earth Sci Inform 2020 Aug 23; 13(4): 1477-84. https://doi.org/10.1007/s12145-020-00505-1
23. Anwar S.M., Majid M., Qayyum A., Awais M., Alnowami M., Khan M.K. Medical Image Analysis using Convolutional Neural Networks: A Review. Journal of Medical Systems 2018 Oct 8; 42(11). https://doi.org/10.1007/s10916-018-1088-1
24. Nogay H, Akinci TC, Yilmaz M. Comparative Experimental Investigation and Application of Five Classic Pre-Trained Deep Convolutional Neural Networks via Transfer Learning for Diagnosis of Breast Cancer. Adv Sci Technol Res J. 2021 Sep 1;15(3):1-8. https://doi.org/10.12913/22998624/137964
25. Szala M, Łatka L, Awtoniuk M, Winnicki M, Michalak M. Neural Modelling of APS Thermal Spray Process Parameters for Optimizing the Hardness, Porosity and Cavitation Erosion Resistance of Al2O3-13 wt% TiO2 Coatings. Processes. 2020 Nov 26;8(12):1544. https://doi.org/10.3390/pr8121544
26. Awtoniuk M, Majerek D, Myziak A, Gajda C. Industrial Application of Deep Neural Network for Aluminum Casting Defect Detection in Case of Unbalanced Dataset. Adv Sci Technol Res J. 2022 Nov 1;16(5):120-8. https://doi.org/10.12913/22998624/154963
27. Gatta GD, Birch WD, Rotiroti N. Reinvestigation of the crystal structure of the zeolite gobbinsite: A single-crystal X-ray diffraction study. Am Mineral. 2010 Mar 25;95(4):481-6. https://doi.org/10.2138/am.2010.3390
28. Machrowska A, KarpińskI R, Maciejewski M, Jonak J, Krakowski P. Application of eemd-dfa algorithms and ann classification for detection of knee osteoarthritis using vibroarthrography. Appl Comput Sci. 2024 Jun 30 https://doi.org/10.35784/acs-2024-18
29. Woźniak M, Połap D, Capizzi G, Sciuto GL, Kośmider L, Frankiewicz K. Small lung nodules detection based on local variance analysis and probabilistic neural network. Comput Methods Programs Biomed. 2018 Jul; 161:173-80. https://doi.org/10.1016/j.cmpb.2018.04.025
30. Ren Y, Zhang L, Suganthan PN. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions. IEEE Comput Intell Mag. 2016 Feb;11(1):41-53. https://doi.org/10.1109/mci.2015.2471235
31. Wang G., Li W., Ourselin S., Vercauteren T., Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, pp. 61-72, Springer International Publishing.
32. Sadoon T.A., Ali M.H. Deep learning model for glioma, meningioma and pituitary classification. International Journal of Advances in Applied Sciences 2021 Mar 1; 10(1): 88. https://doi.org/10.11591/ijaas.v10.i1.pp88-98
33. Dogan A., Birant D. A Weighted Majority Voting Ensemble Approach for Classification. In: 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, Sep 11-15 2019. IEEE.
34. Badža M.M., Barjaktarović M.Č. Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network. Applied Sciences 2020 Mar 15; 10(6): 1999. https://doi.org/10.3390/app10061999
35. Saeedi S., Rezayi S., Keshavarz H., R Niakan Kalhori S. MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. BMC Medical Informatics and Decision Making 2023 Jan 23; 23(1). https://doi.org/10.1186/s12911-023-02114-6
36. Sharma A.K., Nandal A., Dhaka A., Zhou L., Alhudhaif A., Alenezi F., Polat K. Brain tumor classification using the modified ResNet50 model based on transfer learning. Biomedical Signal Processing and Control 2023 Sep; 86: 105299. https://doi.org/10.1016/j.bspc.2023.105299

37. Fooladi S., Farsi H., Mohamadzadeh S. Segmenting the Lesion Area of Brain Tumor using Convolutional Neural Networks and Fuzzy K-Means Clustering. International Journal of Engineering 2023; 36(8): 1556-68. https://doi.org/10.5829/ije.2023.36.08b.15
38. Kang J., Ullah Z., Gwak J. MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers. Sensors 2021 Mar 22; 21(6): 2222. https://doi.org/10.3390/s21062222
39. Jolly K. Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python. Packt Publishing, Limited, 2018.
40. Liang J. Confusion Matrix: Machine Learning. POGIL Activity Clearinghouse 2022; 3(4).
41. Chen, S., Luc, N. M. RRMSE Voting Regressor: A weighting function based improvement to ensemble regression. arXiv preprint arXiv 2022;
42. Kuncheva L. I. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, 2
43. Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Comput Stat. 2020 Jun 13. https://doi.org/10.1007/s00180-020-00999-9
44. Tadeusiewicz R. Sieci neuronowe (Vol. 110). Akademicka Oficyna Wydawnicza RM, 1993.
45. Documentation for Keras library; https://keras.io/api/applications/vgg/ (Accessed: 10.06.2024)
46. Rohith S, Prakash MS, Anitha R, Kumar KS, Yogeswara Sai K. Detection of Brain Tumor using VGG16. In: 2023 8th International Conference on Communication and Electronics Systems (ICCES); 2023 Jun 1-3; Coimbatore, India. IEEE; 2023. https://doi.org/10.1109/icces57224.2023.10192639
47. Rastogi D, Johri P, Tiwari V. Augmentation based detection model for brain tumor using VGG 19. Int J Comput Digit Syst. 2023 May 30; 13(1):1227-37. https://doi.org/10.12785/ijcds/1301100
48. Singamshetty R, Sruthi S, Chandhana K, Kollem S, Prasad CR. Brain Tumor Detection Using the Inception Deep Learning Technique. In: 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC); 2023 Feb 10-11; Mysore, India. IEEE; 2023
49. Azzahra TS, Jessica Jesslyn Cerelia, Farid Azhar Lutfi Nugraha, Anindya Apriliyanti Pravitasari. MRI-Based Brain Tumor Classification Using Inception Resnet V2. Enthusiastic. 2023 Oct 24: 163-75. https://doi.org/10.20885/enthusiastic.vol3.iss2.art4
50. Sujatha K, Rao BS. Densenet201: A Customized DNN Model for Multi-Class Classification and Detection of Tumors Based on Brain MRI Images. In: 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT); 2023 Feb 22-24; Erode, India: IEEE; 2023. https://doi.org/10.1109/icecct56650.2023.10179642
51. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI. IEEE; 2017. https://doi.org/10.1109/cvpr.2017.195
52. Popli R, Kansal I, Verma J, Khullar V, Kumar R, Sharma A. ROAD: Robotics-Assisted Onsite Data Collection and Deep Learning Enabled Robotic Vision System for Identification of Cracks on Diverse Surfaces. Sustainability. 2023 Jun 9;15(12):9314. https://doi.org/10.3390/su15129314
53. Ahmad A., Brown G. Random Projection Random Discretization Ensembles - Ensembles of Linear Multivariate Decision Trees. IEEE Transactions on Knowledge and Data Engineering 2014 May; 26(5): 1225-39. https://doi.org/10.1109/tkde.2013.134
54. Mohammed A., Kora R. A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges. Journal of King Saud University - Computer and Information Sciences 2023 Feb; 35(2): 757-774.
55. Database used during research. https://doi.org/10.6084/m9.figshare.1512427.v5 (Accessed: 10.06.2024