Received: 27 October 2023 | Revised: 24 November 2023 | Accepted: 19 December 2023 | Published online: 27 December 2023

Medinformatics
2024, Vol. 1(1) 10–19
DOI: 10.47852/bonviewMEDIN32021963

RESEARCH ARTICLE

A Weighted Ensemble Approach with Multiple Pre-trained Deep Learning Models for Classification of Stroke

Rusul Ali Jabbar Alhatemi[1] **and Serkan Savaş**[2,*]

[1]*Department of Electronics and Computer Engineering, Çankırı Karatekin University, Türkiye*

[2]*Department of Computer Engineering, Kırıkkale University, Türkiye*

Abstract: Stroke ranks as one of the deadliest diseases globally, emphasizing the crucial need for early diagnosis. This study aims to create a two-stage classification system for stroke and non-stroke images to support early clinical detection. Deep learning (DL), a cornerstone of diagnosis, detection, and prompt treatment, is the primary methodology. Transfer learning adapts successful DL architectures for diverse problems, and ensemble learning combines multiple classifiers for enhanced results. These two techniques are applied to classify stroke using a dataset of stroke and normal images. In the initial stage, six pre-trained models are fine-tuned, with DenseNet, Xception, and EfficientNetB2 emerging as the top performers, achieving validation accuracies of 98.4%, 98.4%, and 98%, respectively. These models serve as base learners within an ensemble framework. A weighted average ensemble method combines them, resulting in a remarkable 99.84% accuracy on a reserved test dataset. This approach exhibits promise for stroke detection, a life-threatening condition, while also demonstrating the effectiveness of ensemble techniques in enhancing model performance.

Keywords: ensemble learning, transfer learning, stroke, biomedical image classification, MR images, deep learning

1. Introduction

Strokes are the third most common cause of disability and the second most common cause of mortality worldwide. Stroke is a major contributor to dementia and depression because it causes the rapid demise of some brain cells from a lack of oxygen when a blocked or torn artery cuts off the blood supply to the brain. Approximately 87% of stroke-related fatalities and years lived with a disability take place in low-income and middle-income nations (Owolabi et al., 2015). Even in certain low-income countries, combining preventative techniques has been successful in lowering stroke mortality.

Current recommendations for the treatment of acute stroke are based on a diagnosis of ischemic stroke established with computed tomography (CT) scanners. Without a way to distinguish between an ischemic and hemorrhagic stroke, doctors in low-resource settings are forced to make tough clinical decisions, such as whether to anticoagulated patients or not and to what degree to regulate their blood pressure. The disparately high rates of stroke in low-income and middle-income countries may be attributable to these patient management issues, insufficient rehabilitation services, a lack of preventive measures, and a lack of knowledge of the potential unique risk factors associated with stroke in these countries (O'Donnell et al., 2010). Understanding the genetic underpinnings of the linkages between risk factors can help guide targeted preventive efforts as part of a larger strategy with four components, including surveillance, prevention, and acute therapy. The evidence base needed to produce the guidelines for stroke prevention, treatment, and rehabilitation in low-income and middle-income countries will be provided by this type of integrated strategy. Taking care of acute stroke in low-resource settings calls for a fresh strategy, one that may revive the original World Health Organization (WHO), worldwide stroke initiative, as a partnership between the WHO the World Federation of Neurology, and the World Stroke Organization to raise awareness of stroke, provide improved surveillance data, and provide guidance for better management and prevention (World Health Organization, 2020).

The assessment of a stroke can be conducted using various imaging techniques, with one notable method utilizing CT imaging, particularly effective for swiftly addressing emergency cases in the initial stages of a stroke. Employing CT images for diagnosing ischemic stroke offers advantages like minimal spatial constraints and rapid image acquisition. Nevertheless, image-based diagnosis presents a considerable challenge, constituting a significant drawback of this approach. Multiple alternative diagnostic methods for strokes, including clinical assessment, magnetic resonance imaging (MRI), and catheter angiography, have been devised. Among these, the CT scan stands out for its capability to deliver prompt results (Jung & Whangbo, 2020). Brain CT scans are frequently employed to assess cerebral conditions; however, promptly and accurately interpreting emergent brain CT images poses a challenge, even for proficient neuroradiologists. The adoption of deep learning (DL) networks in medical image analysis has become commonplace, as they facilitate efficient computer-aided diagnosis (Chen et al., 2022). From this point of view, this

*__Corresponding author:__ Serkan Savaş, Department of Computer Engineering, Kırıkkale University, Türkiye. Email: serkansavas@kku.edu.tr study aims to introduce a new approach called ensemble learning for the classification of strokes from CT images. Ensemble learning methodology is being used for many other disciplines, especially in recent years. Medical image classification and segmentation studies are among these also.

When the studies conducted in the last decade are examined, it has been proven in many studies that convolutional neural network (CNN)-based architectures produce effective results in both segmentation and classification of medical images (Bütüner & Calp, 2022; Calp, 2021; Güngen et al., 2020; Karakış, 2023; Savaş & Damar, 2023). Compared to manual methods, traditional methods, statistical methods, or machine learning (ML) methods, it has been revealed that the CNN algorithm achieves more successful results, thanks to automatic feature extraction from large amounts of data (Buyrukoğlu, 2021; Güler & Polat, 2022). Based on these results, CNN-based architectures have also been used in studies such as segmenting brain lesions or classifying brain diseases. Especially recently, in these studies, instead of producing specific models for each dataset, there has been a tendency to use pre-trained models that have proven their success in different datasets. Since there are different studies related to stroke such as DL, risk factors, carotid artery related, and ML studies, ensemble learning studies, which are the subject of this research and have not been performed at a sufficient level, have been examined in the literature review. However, since transfer learning and fine-tuning studies to be used in ensemble learning constitute the first stage of the proposed approach, transfer learning studies in the literature have also been examined. These studies can be summarized as follows.

Rao et al. (2022) proposed a combination of ResNet50 and a dense layer for the prediction of intracranial hemorrhage on brain images. Their proposal was better than the direct use of ResNet50 pre-trained architecture with 99.6% accuracy, 99.7% specificity, and 99.4% sensitivity results. In another exploration of transfer learning, a study employed multiple CNNs to decipher the electroencephalogram (EEG) of stroke patients, aiming to develop a proficient motor imagery brain–computer interface system (Xu et al., 2021). The models, including EEGNet, DenseNet, Xception, ResNet50, and VGG16, were trained and tested as part of the investigation. The resulting average classification accuracy of the proposed model for 11 subjects was determined to be 66.36%, with EEGNet leading the way.

In their proposal, Jung and Whangbo (2020) introduced a pre-processing algorithm tailored for ischemic stroke, utilizing non-contrast CT data sourced from the middle cerebral artery region (specifically, the NCCT dataset collected from patients). Furthermore, they recommended an adaptive transfer learning algorithm employing the ResNet152 architecture. The study demonstrated a performance improvement of 18.72% compared to other studies evaluated in their research. Chen et al. (2022) proposed a solution for classifying strokes based on findings from unenhanced brain CT images, utilizing a private dataset collected from patients. The classification included categories such as normal, hemorrhage, infarction, and others. The study employed CNN-2, VGG16, and ResNet50 models, varying data sizes, mini-batch sizes, and optimizers. The comparative results indicated that both CNN-2 and ResNet50 outperformed VGG16, achieving an accuracy of 0.9872. However, it is noteworthy that ResNet50 took longer to produce results compared to the other networks.

In another study, Zhang et al. (2021) introduced an intra-domain task-adaptive transfer learning method for stroke detection. Their approach involved refining the model with various binary thresholds of time since stroke (TSS). The methodology incorporated both 2D (ResNet) and 3D CNN (U-Net) architectures, resulting in a Receiver Operating Characteristic-Area Under the Curve value of 0.74. Specifically, for classifying TSS < 4.5 h, their model demonstrated a sensitivity of 0.70 and a specificity of 0.81.

In their proposal, Talo et al. (2019) introduced the ResNet34 model for the classification of normal and abnormal brain MR images. The training of the model involved techniques such as data augmentation, optimal learning rate finder, and fine-tuning. Remarkably, the proposed model demonstrated exceptional performance, achieving a 5-fold classification accuracy of 100% on a dataset consisting of 613 MR images. Dawud et al. (2019) utilized fine-tuning for the brain CT hemorrhage classification task and introduced a novel approach for the AlexNet model incorporating a support vector machine (SVM) classifier. The training and testing of the model involved normal and diseased brain CT images sourced from a hospital in Nigeria. Notably, the model demonstrated a high level of success, achieving an accuracy rate of 93.48% in the study. Cetinoglu et al. (2021) conducted a study to assess the performance of Modified MobileNetV2 and EfficientNetB0 models in detecting and classifying strokes based on diffusion-weighted images. Their custom dataset comprised 1800 slices (900 strokes and 900 normal), distributed into 1400 for training, 200 for validation, and 200 for testing. The study reported a stroke detection accuracy of 96% for the modified MobileNetV2 model and 93% for the EfficientNetB0 model. In the vascular territorial classification of strokes into middle cerebral artery, posterior circulation, or watershed infarction, the modified MobileNetV2 model achieved 93% accuracy, while the modified EfficientNetB0 model achieved 87%. Notably, as literature research did not yield studies using ensemble learning methodology for stroke classification, this study stands out for its promising approach.

The organization of this paper unfolds as follows: The second section elucidates the materials employed in this study and outlines the proposed approach. Following that, the third section presents the results obtained during the course of the study. Subsequently, the fourth section delves into a discussion of the research findings. Finally, the paper is concluded in the fifth section.

2. Materials and Methods

The CNN architecture, widely employed in image recognition and classification studies, is a complex neural network composed of multiple layers, including convolution, pooling, activation, normalization, dropout, fully connected, and classification layers. In the context of image classification, pre-trained models based on the CNN architecture have been extensively analyzed for their performance. The convolution layer, characterized by parameters such as the number of images, image size, and filter size, applies a filter matrix to the input data, treated as a matrix, resulting in convolution and the generation of a new feature map. In the subsequent pooling layer, the feature map's size is reduced through methods like max pooling and average pooling, preparing it for further convolution layers. The CNN model iteratively applies convolution and pooling processes. Following feature extraction and data reduction, the classification layer produces outputs corresponding to the number of objects targeted for classification. This sequential process forms the foundation of CNN-based image recognition systems (Çınarer et al., 2020; Güler & Yücedağ, 2022).

2.1. Research design

The classification process for predicting stroke or non-stroke in the brain is presented in this section (Figure 1). This process starts with data acquisition. After loading the data, we implemented some data augmentation techniques since DL studies use a better with much amount of data. Then, the data pre-processing step is finished by splitting the data into three parts, train, validation, and test. Next, pre-trained Deep Neural Network are applied to the training and validation data to define the best three models among them. For the application process, the models are fine-tuned appropriately to the problem. The first step of the proposed approach is completed with the results of the pre-trained DNNs. In the second step, the top-3 models of the first step are defined as base learners of the ensemble learning process. We brought these three models together to get a weighted average prediction result. Finally, the ensemble models give a classification result for the test data. This result is the average of scores that resulted from the best three models.

2.2. Data

To employ the suggested transfer learning-based models, the brain MR images dataset is divided into two classes: those with and those without strokes. The MR images are first subjected to many pre-processing procedures for image augmentation and enhancement. The initial dataset contains 2376 MR images. There are 1426 and 950 images in non-stroke and stroke classes, respectively. The Kaggle website offers access to the MRI scans and study dataset is get from the Kaggle. Since the data used in the study were removed from the Kaggle platform after the study was conducted, no data link is provided. However, the study data can be sent to researchers who request it, with privacy/ethical restrictions. Kaggle is a platform for data science, ML, and DL that facilitates collaboration and competition among a global community of data scientists, researchers, and practitioners. Founded in 2010, Kaggle provides a platform where users can access and share datasets, explore, and build models in a cloud-based environment, and participate in ML competitions to solve real-world problems. The platform offers a diverse range of datasets and challenges, enabling participants to hone their skills, learn from others, and contribute to the advancement of data science. Kaggle (2023) also serves as a hub for knowledge exchange through forums, notebooks, and discussions, fostering a vibrant ecosystem for the exploration and application of data-driven solutions. A larger MR training dataset has been produced by expanding this collection. Six pre-trained models are used to train and assess the transfer learning model.

After data acquisition, the initial stage is data augmentation, to increase the number of MR images for training the pre-trained models. The original dataset is supplemented with a variety of random alterations (rotations, brightness changes, etc.,) using the image data generator function. This process serves as a regulator and lowers overfitting during DL model training. Data augmentation has a close relationship to oversampling in data analysis.

The second goal of the preparation stage is resizing the dataset images to 224 × 224 × 3 so that they can be normalized and made consistent with the used pre-trained models. The process of separating the study dataset into three parts for training, validation, and testing is the last step in the preparation process. Seventy percentage of the overall dataset is stored for training. The remaining images were stored for model validation (20%) and testing (10%). To reduce the sample difference between the classes in the study and to prevent the formation of a dominant class, the images were rotated by 45° to reach 2500 images. Two hundred and fifty of these images were used for testing, while the remaining images were used in the training and validation phase.

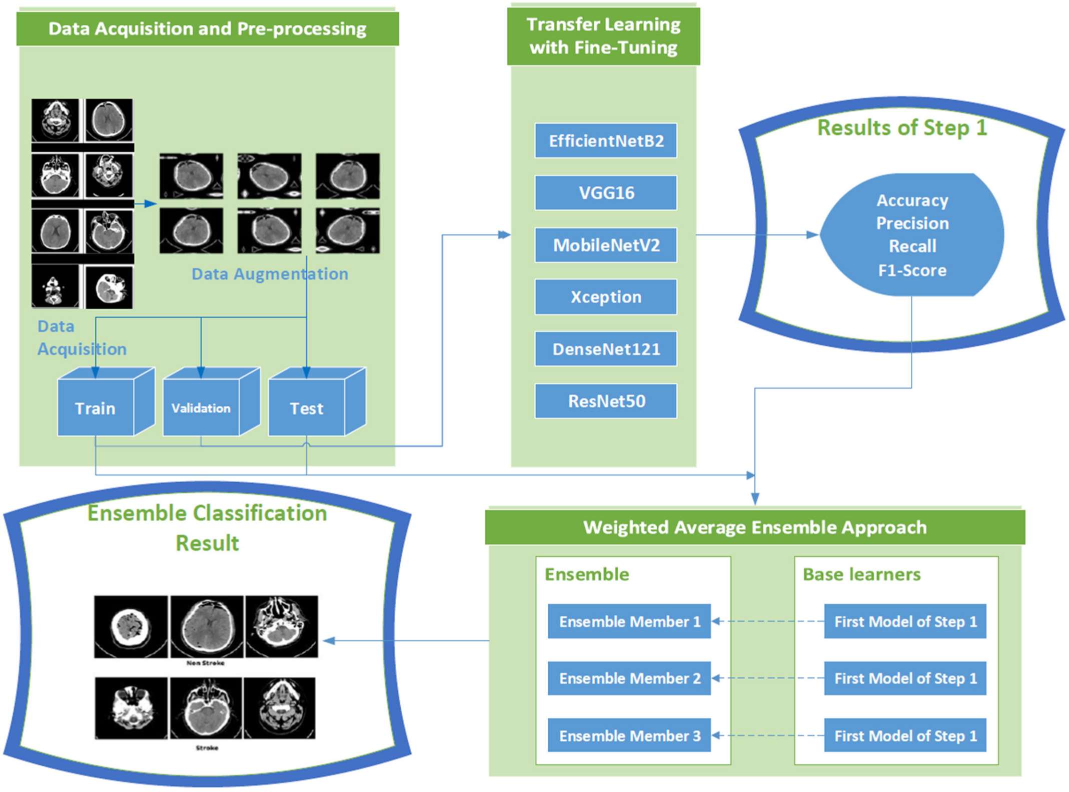

Figure 1. Block diagram of the approach used in the study

2.3. Transfer learning

Transfer learning is the process of resolving a separate but related problem utilizing the relevant parts of a pre-trained DL model. Often, programmers select the model's components that need to be re-trained and applied to the new task. In our research, we used six pre-trained CNN architectures, which are successful for many other problems (ResNet50, MobileNetV2, EfficientNetB2, VGG16, Xception, and DenseNet121).

The ResNet50, which is a widely used ResNet model, consists of 48 convolution layers, 1 MaxPool layer, and 1 average pool layer. The number of floating-point operations is 3.8 *x* 109 (He et al., 2016). MobileNetV2 is an architecture, which uses depthwise separable convolution that is suited for mobile devices or other devices with limited processing capabilities. In this model, an enhanced module with an inverted residual structure is included. Non-linearities in thin layers are removed this time. Using this architecture as the basis for feature extraction, contemporary results are also obtained for item detection and semantic segmentation (Sandler et al., 2018). In contrast to conventional methods, the EfficientNet scaling strategy uses a set of pre-determined scaling coefficients to uniformly scale network width, depth, and resolution. To scale these parameters of the network, EfficientNet employs a compound coefficient (Tan & Le, 2019). In the VGG16 architecture, the image is sent to the network with (224 × 224 × 3) sizes. The feature extraction layers of the structure continue, with two convolutional layers with 64 filters, two convolutional layers with 128 filters, three convolutional layers with 256 filters, three convolutional layers with 512 filters, and three convolutional layers with 256 filters. Between all these convolutional layer blocks, there is a pooling layer that uses the max-pooling technique (Simonyan & Zisserman, 2014). An Xception is a type of CNN that uses dense blocks to connect all layers (with matching feature-map sizes) directly with one another, creating dense connections between the levels. Each layer sends its feature and receives additional inputs from all preceding levels (Chollet, 2016). DenseNet121 has one 7 × 7 filter, fifty-eight 3 × 3 filters, sixty-one 1 × 1 convolution layers, and four AvgPool layers in its structure. All layers, that is, those within the same dense block and transition layers, spread their weight across multiple inputs, allowing deeper layers to use previously extracted features (Ruiz, 2018).

Table 1 summarizes the important hyperparameters that are employed throughout the execution of the pre-trained models. Each of the pre-trained models used in the study was fine-tuned with the parameters specified in Table 1. After the training process, the final weights of the models were saved for use in ensemble learning.

To incorporate the pre-trained models into the application using the transfer learning method, the fine-tuning adjustments used in Table 1 were made. Each model retrieved from the Keras applications library was applied to the problem used in the study using these parameters. Since the number of classes in the ImageNet competition is 1000, the models retrieved from the Keras library were trained for this number of outputs. While the weights of the models included in this study were the weights in the ImageNet competition, in the fully connected layers, also called classification layers, the GlobalAveragePooling method was first applied, and then the model output was reduced to 2 classes (stroke and non-stroke), which is the classification problem. In addition, a callback for early stopping over the "loss" value was also added during the training of the study.

The tests in the study were performed using the Python programming language in the Google Colab environment, a platform that supports online ML and DL studies. TensorFlow (version 2.10.0), Matplotlib (version 3.6.0), and Sklearn (version 1.2.0) libraries were used to perform the analysis.

Table 1. Fine-Tuning of hyperparameters

No	Parameter	Values
1	Optimizer	Adam, beta_1 = 0.9, beta_2 = 0.999
2	Learning rate	0.001 to min 0.0000001
3	Loss function	categorical_crossentropy
4	Metrics	Accuracy, precision, recall, F1-score
5	Batch size	32
6	Epochs	50
7	Patience	5

2.4. Ensemble learning

Ensemble learning involves the amalgamation of diverse learning algorithms to collectively improve performance or enhance existing models by integrating multiple models into a reliable single model (Cao et al., 2020). While DL models have demonstrated proficiency in various applications, there remains an opportunity to employ a combination of DL models using ensemble techniques to achieve comparable objectives. The study utilizes a randomized weighted ensemble, an ensemble technique that assigns weights to predictions from each ensemble member. These weights are combined to calculate an aggregated prediction, as illustrated in Equation (1). The optimization of weights involves a search process using randomized search, based on the Dirichlet distribution and conducted on a validation dataset (Borges, 2019).

$$w_1.[\hat{y}_1] + w_2.[\hat{y}_2] + \cdots w_n.[\hat{y}_n] = [\hat{y}] \tag{1}$$

where w is the weight of each member, $\hat{y}$ is the output of each member, and $\hat{y}$ is the weighted average ensemble output. The deepstack module for ensemble learning was used in the study and version 0.0.9 of this library was included in the program.

2.5. Evaluation metrics

Many different evaluation measures can be used to test a model. These include classification accuracy, loss, and confusion matrices, F1-score, recall, precision, sensitivity, specificity, etc. Typically, we refer to classification accuracy when we use the word accuracy. There are four categorizations used to compute these metrics from Equations (2)–(5) as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \tag{4}$$

$$F1Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \tag{5}$$

In the Equations;

- A TP result is one in which the model accurately predicts the positive class.

- A TN result is one in which the model accurately predicts the negative class.
- An FP result occurs when the model guesses the positive class falsely.
- An FN result occurs when the model guesses the negative class falsely.

Equation (2) defines accuracy as the total successfully categorized examples divided by the total examples that were classified. Equation (3) defines precision as a rate of correctly predicted samples for the positive class. For a classification problem with two classes, recall (also called sensitivity) is calculated in Equation (4) as the ratio of true positives to both true positives and false negatives. The F1-score is calculated as in Equation (5) and is the harmonic mean of precision and recall (Nighania, 2018).

The loss function serves as a crucial metric, assessing the disparity between the predicted value and the actual label. In the context of image processing, the loss layer of a neural network contrasts the network's output with the ground truth. The study adopts the categorical cross-entropy as the chosen loss function, calculated as outlined in Equation (6) (Zhao et al., 2017).

$$Loss = -\sum_{i=1}^{\substack{Output\\Size}} y_i \log \hat{y}_i \tag{6}$$

3. Results

3.1. Results of the transfer learning models

In this section, the results of the transfer learning model according to all metrics used like accuracy, loss, precision, recall, and F1-score are explained. The accuracy graphs are given in Figure 2 (ResNet50, MobileNetV2, EfficientNetB2, VGG16, Xception, and DenseNet121).

As seen in Figure 2, the validation average accuracy rates of the models are 96.8%, 96%, 98%, 57.9%, 98.4%, and 98.4% for the ResNet50, MobileNetV2, EfficientNetB2, VGG16, Xception, and DenseNet121, respectively. Although the number of epochs was defined as 50, during the training of ResNet50 and Xception models, early stopping call-back was activated with the function it was defined and stopped the training at the 30th epoch.

Among the models, the VGG16 model was significantly the least accurate, as shown in Figure 2(d). The training accuracy of this model was 60.34% while the validation accuracy was 57.9%. The test accuracy results are explained in Figure 4 together with the confusion matrix. Among the other models, EfficientNetB2, Xception, and DenseNet121 were the three models that exceeded 98% accuracy. Loss graphs of the models were also obtained for a more comprehensive evaluation of the performance of these models. The loss graphs of the ResNet50, MobileNetV2, EfficientNetB2, VGG16, Xception, and DenseNet121 models are given in Figure 3, respectively.

The important point in the Loss parameter is that the Loss value of the model converges to 0. Thus, it is concluded that the confusion rate of the prediction result between the labeled data of the model decreases. The Loss values of all models converged to zero as seen in Figure 3. The validation Loss values of the models are 23×10^{-5}, 149×10^{-5}, 6×10^{-5}, 68160×10^{-5}, 95×10^{-5}, and $69 \times 10^{-5}$for the ResNet50, MobileNetV2, EfficientNetB2, VGG16, Xception, and DenseNet121, respectively. Figure 3 shows that the Loss values of MobileNetV2 and EfficientNetB2 models oscillate up to the first 30 epochs, while the other models stabilize after the first 10 epochs. This means that the confusion rate between the labeled data classes of the models is almost negligible. When we look

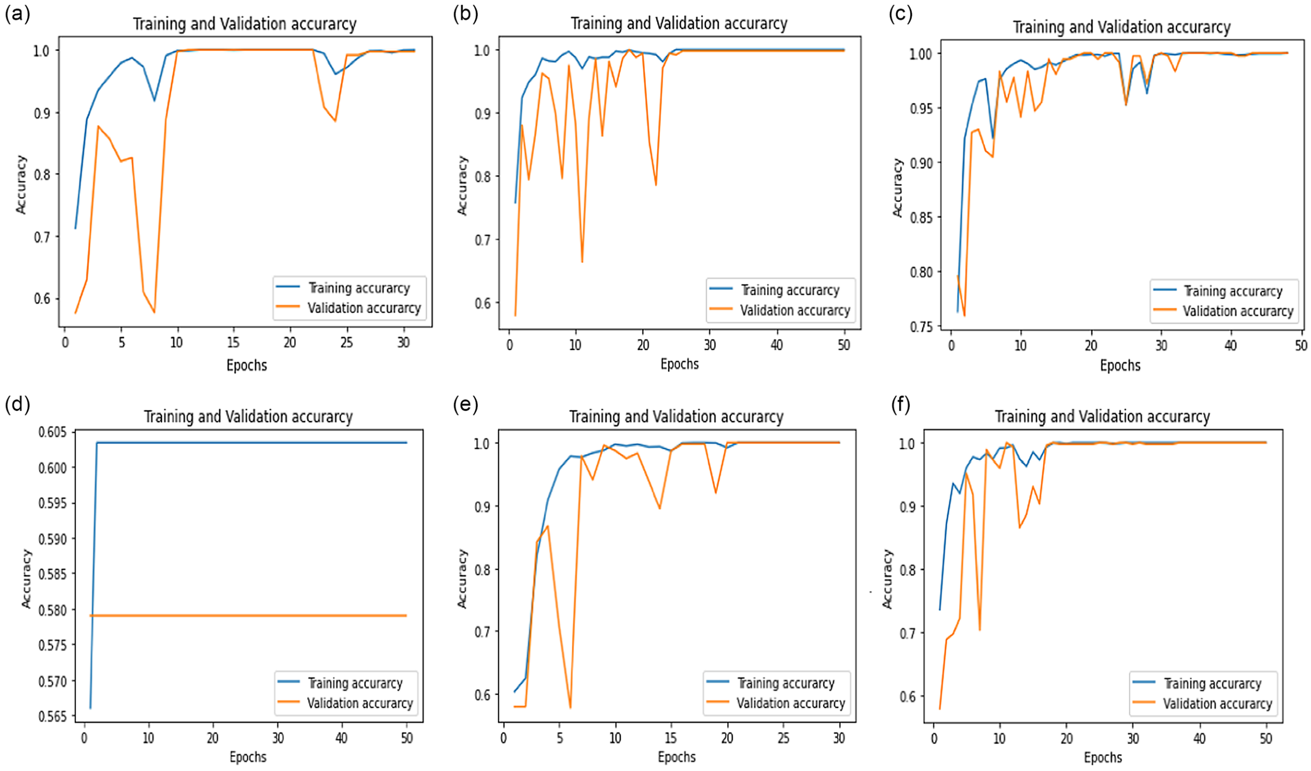

Figure 2. Accuracy results of (a) ResNet50, (b) MobileNetV2, (c) EfficientNetB2, (d) VGG16, (e) Xception, and (f) DenseNet121 models

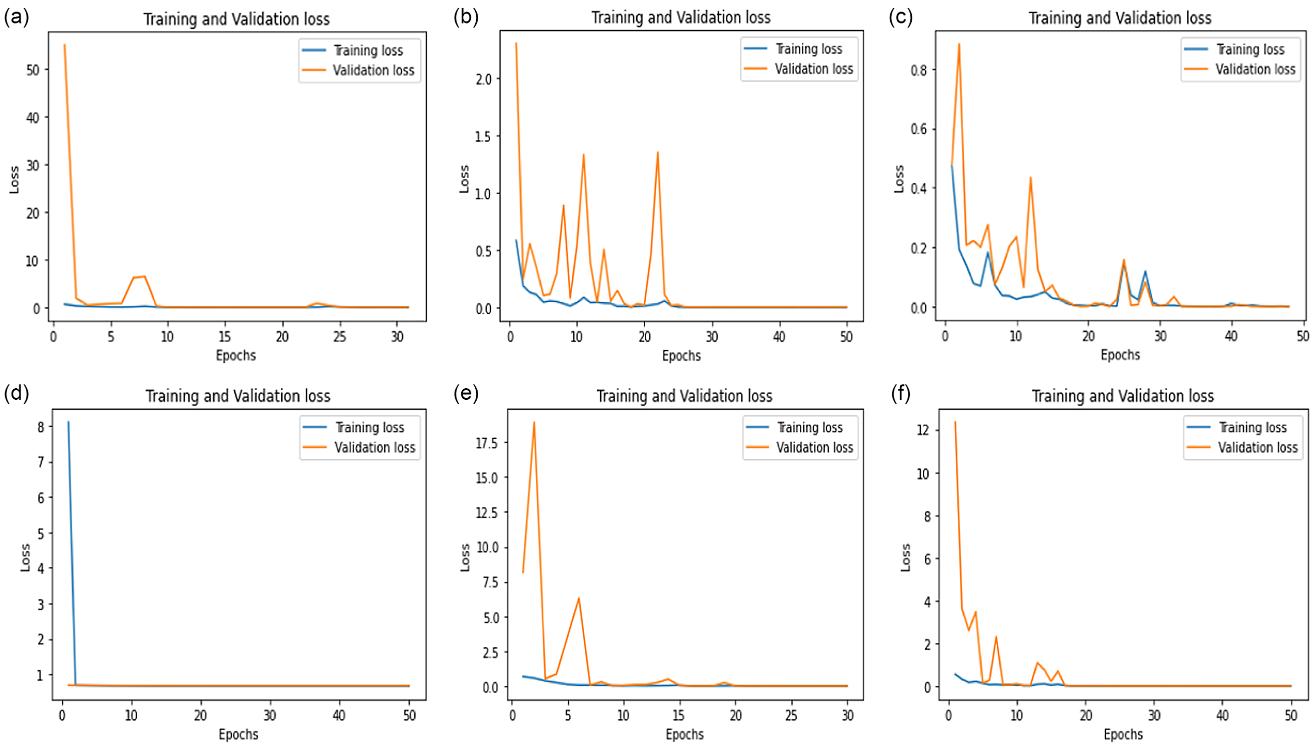

Figure 3. Loss results of (a) ResNet50, (b) MobileNetV2, (c) EfficientNetB2, (d) VGG16, (e) Xception, and (f) DenseNet121 models

at the results of the Loss values of the models, although all models obtained very low Loss values, the VGG16 model reached the highest Loss rate among the models.

After evaluating the performance of the models on the training and validation data, their performance on the test data was finally evaluated to determine the three most successful models in the first stage. These performances were realized through confusion matrix metrics. The confusion matrix is an important criterion for evaluating the performance of DL models. It can be used especially in health studies to see the correct and incorrectly predicted values for each class and to measure the ratios. The confusion matrix results of the pre-trained models used in the study are given in Figure 4.

As seen in Figure 4, there are 250 images in the test dataset consisting of 125 images from both classes. The ResNet50 model classified all non-stroke images truly but 8 of 125 images in the stroke class are predicted as non-stroke, which is a false prediction. The MobileNetV2 model also predicted all non-stroke images truly, but the MobileNetV2 made 10 false predictions for the stroke class, which is worse than ResNet50. The EfficientNetB2 model predicted truly 124 of 125 images for the non-stroke and 121 of 125 images for the stroke classes. EfficientNetB2 model prediction for 4 images from the stroke class and 1 image from the non-stroke class was false. As seen in Figure 4(d), the VGG16 model predicted all samples as in the stroke like a single class. This confusion matrix result is generally encountered as the output of failed models. It is an important indicator that the model fails on test data. These results of the VGG16 model support the training and validation accuracy and loss rates. As the most unsuccessful model among all pre-trained models, it is concluded that VGG16 is not suitable for the problem of this study. The Xception model only made 4 false predictions for the stroke class and the DenseNet121 model made 3 and 1 false predictions for the stroke and non-stroke classes, respectively.

After the fine-tuning of the pre-trained models, the training and validation evaluation, and after obtaining the confusion matrices on the test data, confusion matrix metrics were also generated for each model. These obtained ratios are presented in Table 2.

As seen in Table 2, for the ResNet50 model, precision, recall, and F1-score of the stroke class are 100%, 93.6%, and 96.7%, respectively. For the non-stroke class, these are 93.98%, 100%, and 96.9%, respectively. The stoke class precision, recall, and F1-score of the MobileNetV2 are 100%, 92%, and 95.83%, respectively. The non-stroke class rates of this model are 92.6%, 100%, and 96.15%, respectively. EfficientNetB2 model's precision, recall, and F1-score rates are 99.18%, 96.8%, and 97.98%, respectively, and for the other class, these metrics are 96.88%, 99.2%, and 98.02%, respectively. VGG16 model's results are not successful for this problem. The precision, recall, and F1-score of the Xception model for the stroke class are 100%, 96.8%, and 98.37%, respectively. For the non-stroke class, these metrics are 96.9%, 100%, and 98.43%, respectively. Lastly, DesNet121 metrics for the stoke class, like precision, recall, and F1-score are 99.17%, 97.6%, and 98.39%, respectively. For the non-stroke class, they are 97.64%, 99.2%, and 98.4%, respectively.

After all these evaluations, DenseNet121, Xception, and EfficientNetB2 were determined as the three most successful models as a result of the first step of the proposed approach in the study. These models were used as base learners in the ensemble learning phase.

3.2. Results of ensemble learning

The best models used in this process are the DenseNet121, Xception, and EfficientNet models. Table 3 represents the scores

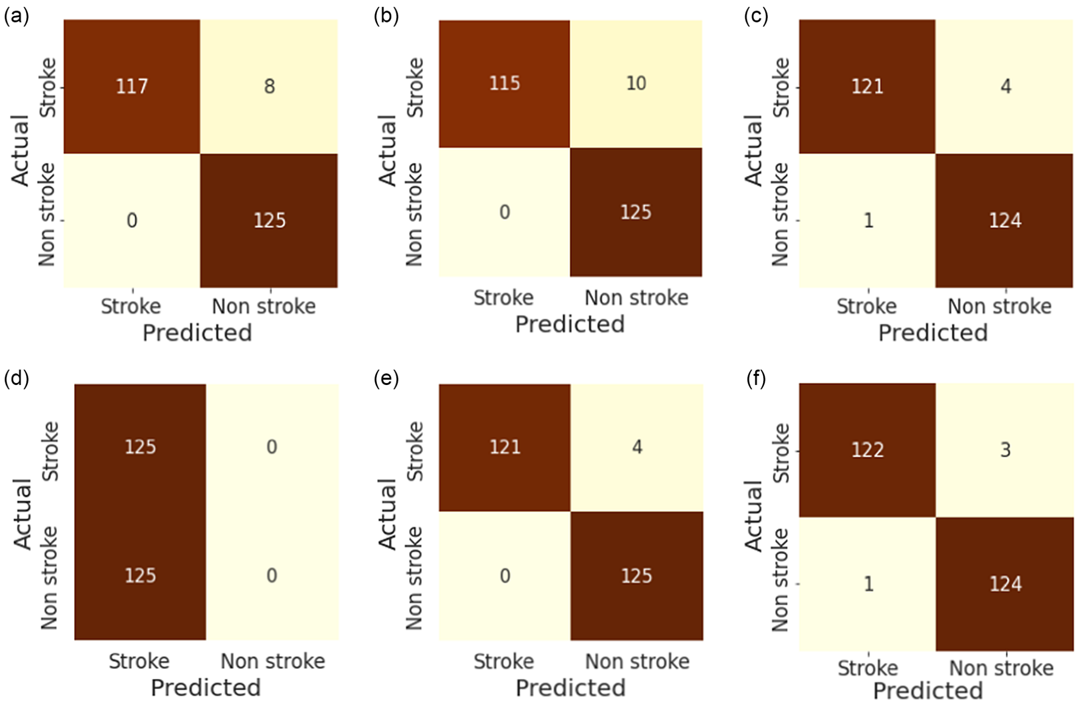

Figure 4. Confusion matrices of (a) ResNet50, (b) MobileNetV2, (c) EfficientNetB2, (d) VGG16, (e) Xception, and (f) DenseNet121 models

Table 2. Precision, recall, and F1-score results for the models for each class

	Class	Precision	Recall	F1-score	Support
ResNet50	Stroke	1.00	0.936	0.967	125
	Non-stroke	0.9398	1.00	0.969	125
MobileNetV2	Stroke	1.00	0.92	0.9583	125
	Non-stroke	0.926	1.00	0.9615	125
EfficientNetB2	Stroke	0.9918	0.9680	0.9798	125
	Non-stroke	0.9688	0.9920	0.9802	125
VGG16	Stroke	0.5000	1.0000	0.6667	125
	Non-stroke	0.0000	0.0000	0.0000	125
Xception	Stroke	1.00	0.968	0.9837	125
	Non-stroke	0.969	1.00	0.9843	125
DenseNet121	Stroke	0.9919	0.976	0.984	125
	Non-stroke	0.9764	0.992	0.984	125

Table 3. Ensemble learning results

Model	Results
EfficientNetB2	– Weight: 0.0230 accuracy_score: 0.8703
DenseNet121	– Weight: 0.0002 accuracy_score: 0.9335
Xception	– Weight: 0.9768 accuracy_score: 0.9963
Ensemble model	99.84%

of the ensemble learning using the best three models. It represents the score for every pre-trained model during the application of the ensemble learning and the total score.

The models used in the ensemble model structure were taken from the Keras application library in the transfer learning phase and adapted to the research problem through fine-tuning. These models were trained according to the classification structure of the research problem in the training phase and their weights were recorded. These weights were used as a base learner for ensemble learning in the deepstack structure. Within this structure, each model contributed to the ensemble learning process with different weights according to the Dirichlet distribution formulation.

The ensemble model has a 99.84% accuracy score because it combines the advantages of the best models represented. The EfficientNetB2 model has an accuracy rate equal to 87.03% with a weight rate of 0.023, the DenseNet121 model has an accuracy rate equal to 93.35% with a weight rate of 0.0002, and the accuracy rate of the Xception model is 99.63% with a rate of 0.9768, as a single base learner in the deepstack structure. When we combined all these models with the weighted average technique in ensemble learning, the accuracy score for the test dataset became 99.84%, which is a perfect score for the problem.

4. Discussion

The outcomes of the proposed approach, including transfer learning and ensemble learning models, are contrasted in Table 4. Table 4 represents a comparison between the proposed approach with the related works according to models used, aims, and results. Stroke is one of the most important health problems, which ranks second worldwide among the diseases that result in death. Therefore, studies on this subject are also of great importance. Recent studies using transfer learning and ensemble learning techniques for classification and detection processes are listed in Table 4. More than 90% success was achieved in most of these studies.

In their study, Dawud et al. (2019) employed an AlexNet combined with a SVM for hemorrhage classification, achieving an accuracy of 93.48%. Talo et al. (2019) utilized ResNet34 to successfully classify normal and abnormal brain MR images, achieving a perfect accuracy of 100%. Xu et al. (2021) explored the effectiveness of various models, including EEGNet, DenseNet, Xception, ResNet50, and VGG16, in designing a motor imagery brain–computer interface system, achieving an accuracy of 66.36%. Zhang et al. (2021) focused on stroke detection, utilizing ResNet and U-Net models and reporting a specificity result of 0.81. Cetinoglu et al. (2021) employed MobileNetV2 and EfficientNetB0 for the detection and vascular territorial classification of stroke, achieving accuracy rates of 96% and 93%, respectively. In the study conducted by Rao et al. (2022), ResNet50 was employed for predicting intracranial hemorrhage, resulting in an impressive accuracy of 99.6%. Lastly, Chen et al. (2022) utilized CNN-2, VGG16, and ResNet50 for the classification of stroke, achieving a high accuracy of 98.72%. These studies collectively demonstrate the diverse application of CNN models in medical image analysis, showcasing their effectiveness in tasks ranging from hemorrhage classification to stroke detection and intracranial hemorrhage prediction.

The conclusion drawn from this is that transfer learning and ensemble learning techniques can be promising results for stroke. However, it is seen that the ensemble learning approach proposed in the study is superior to previous studies. In addition, it is seen that the models used alone increase the performance. From this point of view, an approach that can lead to future studies has been put forward.

Ensemble learning is widely utilized in ML and DL to enhance the performance and robustness of predictive models by combining multiple individual models. The fundamental principle behind ensemble methods is that aggregating diverse models can yield improved overall predictions compared to a single model. In the field of ML, ensemble learning encompasses various approaches such as bagging, boosting, and stacking, which have been extensively studied. Bagging techniques, such as random forests, involve training multiple models on different subsets of the training data and then combining their predictions through voting or averaging. Boosting methods, such as AdaBoost and gradient boosting, iteratively train weak models by assigning higher weights to misclassified instances, thereby creating a stronger final model. Stacking, on the other hand, combines predictions from multiple models using another model, commonly referred to as a meta-learner, to make the final prediction (Polikar, 2006).

Ensemble learning has also garnered significant attention within the DL community. In DL, ensemble approaches focus on combining predictions from multiple DNNs to enhance model performance. One commonly employed technique is model averaging, where multiple neural networks with different initializations or architectures are independently trained, and their predictions are averaged to obtain the final prediction. Another approach is known as model stacking, where the outputs of multiple neural networks serve as input features to a meta-learner that learns to make the final prediction. Ensemble learning in DL has demonstrated the ability to improve model generalization, mitigate overfitting, and enhance prediction accuracy across various domains, including computer vision, natural language processing, and speech recognition (Deng & Yu, 2014).

Ensemble learning techniques are particularly effective when the individual models within the ensemble possess diverse strengths and weaknesses. By combining multiple models, ensemble learning can capture a broader range of patterns and enhance the model's resilience to noise or outliers in the data. Furthermore, ensemble learning provides a means to explore different hypotheses and model architectures, enabling more reliable and confident predictions. However, it is important to note that ensemble learning may introduce additional computational and resource requirements due to training and combining multiple models. Therefore, the selection and design of ensemble learning methods should be carefully considered based on the specific task, dataset, and available resources (Brown, 2011).

Table 4. Comparison of proposed work with related work

Author	Model	Aim	Result (Acc)
Dawud et al. (2019)	AlexNet+SVM	Hemorrhage classification	93.48%
Talo et al. (2019)	ResNet34	Classify normal and abnormal brain MR images	100%
Xu et al. (2021)	EEGNet, DenseNet, Xception, ResNet50, and VGG16	To design an effective motor imagery brain–computer interface system	66.36%
Zhang et al. (2021)	ResNet, U-Net	Stroke detection	0.81 (Specificity)
Cetinoglu et al. (2021)	MobileNetV2 and EfficientNetB0	Detection and vascular territorial classification of stroke	96% and 93%
Rao et al. (2022)	ResNet50	Prediction of intracranial hemorrhage	99.6%
Chen et al. (2022)	CNN-2, VGG16, and ResNet50	Classification of stroke	98.72%
Proposed work	DenseNet121	Classification of stroke	98.4%
	Xception		98.4%
	EfficientNetB2		98%
	Ensemble model		99.84%

5. Conclusion

An efficient approach for classifying stroke or non-stroke automatically from MRI data was investigated in this work. The approach is built on transfer learning, which utilizes well-known CNN architectures and ensemble learning. Stroke medical image data analysis has always been difficult, and the most difficult part is choosing the distinguishing characteristics with some understandable indicators.

The experimental results demonstrated that, despite the short amount of the given dataset, a classifier could identify stroke with pre-trained DL models. The DenseNet121 and Xception DL models outperformed the state of the art with an accuracy of 98.4%, demonstrating their supremacy over the other models. Later, the total score of implementing ensemble learning using the best three models is obtained as 99.84%, using EfficientNetB2, DenseNet121, and Xception. Even while diagnosing stroke disorders with MR images has seen considerable progress recently, it is still a long way from meeting the criteria for a clinical diagnosis. When creating any predictive models for clinical diagnosis, generalization, and interpretability should be taken into consideration. Most current approaches can only categorize one condition versus a healthy control, thus an intelligent system that can identify many disorders would be greatly desired. In this context, the contributions of this study can be listed as follows.

- Transfer Learning and Ensemble Approach: The study contributes by exploring an efficient approach for automated stroke classification from MRI data. Leveraging transfer learning and ensemble learning techniques, the research employs well-established CNN architectures, demonstrating the efficacy of these methods in handling the challenging task of stroke detection in medical images.
- Model Performance: The experimental results highlight the success of the proposed classifier, even with a limited dataset. Notably, the DenseNet121 and Xception DL models exhibit superior performance with an accuracy of 98.4%, surpassing current state-of-the-art methods. The implementation of ensemble learning further enhances accuracy, reaching a commendable 99.84% using EfficientNetB2, DenseNet121, and Xception, showcasing the robustness of the approach.
- Challenges and Considerations for Clinical Diagnosis: The conclusion underscores the persisting challenges in stroke medical image analysis, particularly in achieving clinical diagnosis standards. The need for generalization and interpretability in predictive models for clinical use is emphasized. The study prompts a thoughtful consideration of creating intelligent systems capable of identifying various disorders, going beyond binary categorizations of a single condition versus a healthy control.
- Future Directions: The research suggests intriguing possibilities for future work, such as enhancing findings by integrating MR images with other data sources like electronic medical records, EEG, and structural MR images. The exploration of comprehensive datasets and the utilization of multiple DL models are proposed as avenues to develop specialized models capable of identifying different types of brain strokes, indicating a potential direction for advancing the field.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Since the data used in the study were removed from the Kaggle platform after the study was conducted, no data link is provided. However, the study data can be sent to researchers who request it, with privacy and ethical restrictions.

References

Borges, J. (2019). *DeepStack: Ensembling Keras deep learning models into the next performance level*. Retrieved from: https://github.com/jcborges/DeepStack

Brown, G. (2011). Ensemble learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 312–320). Germany: Springer. https://doi.org/10.1007/978-0-387-30164-8_252

Bütüner, R., & Calp, M. H. (2022). Diagnosis and detection of COVID-19 from lung tomography images using deep learning and machine learning methods. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(2), 190–200. https://ijisae.org/index.php/IJISAE/article/view/1843

Buyrukoğlu, S. (2021). Early detection of Alzheimer's disease using data mining: Comparison of ensemble feature selection approaches. *Konya Journal of Engineering Sciences*, *9*(1), 50–61. https://doi.org/10.36306/konjes.731624

Calp, M. H. (2021). Use of deep learning approaches in cancer diagnosis. In U. Kose & J. Alzubi (Eds.), *Deep learning for cancer diagnosis* (pp. 249–267). Germany: Springer. https://doi.org/10.1007/978-981-15-6321-8_15

Cao, Y., Geddes, T. A., Yang, J. Y. H., & Yang, P. (2020). Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, *2*(9), 500–508. https://doi.org/10.1038/s42256-020-0217-y

Cetinoglu, Y. K., Koska, I. O., Uluc, M. E., & Gelal, M. F. (2021). Detection and vascular territorial classification of stroke on diffusion-weighted MRI by deep learning. *European Journal of Radiology*, *145*, 110050. https://doi.org/10.1016/j.ejrad.2021.110050

Chen, Y. T., Chen, Y. L., Chen, Y. Y., Huang, Y. T., Wong, H. F., Yan, J. L., & Wang, J. J. (2022). Deep learning–based brain computed tomography image classification with hyperparameter optimization through transfer learning for stroke. *Diagnostics*, *12*(4), 807. https://doi.org/10.3390/diagnostics12040807

Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *arXiv Preprint: 1610.02357*.

Çınarer, G., Emiroğlu, B. G., Arslan, R. S., & Yurttakal, A. H. (2020). Brain tumor classification using deep neural network. *Advances in Science, Technology and Engineering Systems Journal*, *5*(5), 765–769. https://doi.org/10.25046/aj050593

Dawud, A. M., Yurtkan, K., & Oztoprak, H. (2019). Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. *Computational Intelligence and Neuroscience*, *2019*, 4629859. https://doi.org/10.1155/2019/4629859

Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Güler, O., & Polat, K. (2022). Classification performance of deep transfer learning methods for pneumonia detection from chest

X-Ray images. *Journal of Artificial Intelligence and Systems*, *4*(1), 107–126. https://doi.org/10.33969/AIS.2022040107

Güler, O., & Yücedağ, İ. (2022). Hand gesture recognition from 2D images by using convolutional capsule neural networks. *Arabian Journal for Science and Engineering*, *47*(2), 1211–1225. https://doi.org/10.1007/s13369-021-05867-2

Güngen, C., Polat, Ö., & Karakış, R. (2020). Classification of brain tumors using convolutional neural network from MR images. In *28th Signal Processing and Communications Applications Conference,* 1–4. https://doi.org/10.1109/SIU49456.2020.9302090

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Jung, S.-M., & Whangbo, T.-K. (2020). A deep learning system for diagnosing ischemic stroke by applying adaptive transfer learning. *Journal of Internet Technology*, *21*(7), 1957–1968. https://jit.ndhu.edu.tw/article/view/2431

Kaggle. (2023). *Level up with the largest AI & ML community.* Retrieved from: https://www.kaggle.com

Karakış, R. (2023). MI-STEG: A medical image steganalysis framework based on ensemble deep learning. *Computers, Materials & Continua*, *74*(3), 4649–4666. https://doi.org/10.32604/CMC.2023.035881

Nighania, K. (2018). *Various ways to evaluate a machine learning model's performance*. Retrieved from: https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

O'donnell, M. J., Xavier, D., Liu, L., Zhang, H., Chin, S. L., Rao-Melacini, P., . . ., & Yusuf, S. (2010). Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): A case-control study. *The Lancet*, *376*(9735), 112–123. https://doi.org/10.1016/S0140-6736(10)60834-3

Owolabi, M. O., Akarolo-Anthony, S., Akinyemi, R., Arnett, D., Gebregziabher, M., Jenkins, C., . . ., & Ovbiagele, B. (2015). The burden of stroke in Africa: A glance at the present and a glimpse into the future. *Cardiovascular Journal of Africa*, *26*(2), S27–S38. https://doi.org/10.5830%2FCVJA-2015-038

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Rao, B. N., Mohanty, S., Sen, K., Acharya, U. R., Cheong, K. H., & Sabut, S. (2022). Deep transfer learning for automatic prediction of hemorrhagic stroke on CT images. *Computational and Mathematical Methods in Medicine*, *2022*, 3560507. https://doi.org/10.1155/2022/3560507

Ruiz, P. (2018). *Understanding and visualizing DenseNets*. Retrieved from: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Savaş, S., & Damar, Ç. (2023). Transfer-learning-based classification of pathological brain magnetic resonance images. *ETRI Journal*, 1–14. https://doi.org/10.4218/etrij.2022-0088

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint: 1409.1556*.

Talo, M., Baloglu, U. B., Yıldırım, Ö., & Rajendra Acharya, U. (2019). Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research*, *54*, 176–188. https://doi.org/10.1016/j.cogsys.2018.12.007

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, *97*, 6105–6114. Retrieved from: https://proceedings.mlr.press/v97/tan19a.html

World Health Organization. (2020). *WHO package of essential noncommunicable (PEN) disease intervention for primary health care.* Retrieved from: https://www.who.int/publications/i/item/9789240009226

Xu, F., Miao, Y., Sun, Y., Guo, D., Xu, J., Wang, Y., . . ., & Zhang, Y. (2021). A transfer learning framework based on motor imagery rehabilitation for stroke. *Scientific Reports*, *11*(1), 19783. https://doi.org/10.1038/s41598-021-99114-1

Zhang, H., Polson, J. S., Nael, K., Salamon, N., Yoo, B., El-Saden, S., . . ., & Arnold, C. W. (2021). Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time. *Computerized Medical Imaging and Graphics*, *90*, 101926. https://doi.org/10.1016/j.compmedimag.2021.101926

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, *3*(1), 47–57. https://doi.org/10.1109/TCI.2016.2644865

How to Cite: Alhatemi, R. A. J., & Savaş, S. (2024). A Weighted Ensemble Approach with Multiple Pre-trained Deep Learning Models for Classification of Stroke. *Medinformatics*, *1*(1), 10–19. https://doi.org/10.47852/bonviewMEDIN32021963