IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 December 2022, accepted 10 January 2023, date of publication 16 January 2023, date of current version 23 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3237542

RESEARCH ARTICLE

SSCLNet: A Self-Supervised Contrastive Loss-Based Pre-Trained Network for Brain MRI Classification

ANIMESH MISHRA, RITESH JHA, AND VANDANA BHATTACHARJEE
Birla Institute of Technology, Mesra, Ranchi 835215, India

Corresponding author: Vandana Bhattacharjee (vbhattacharya@bitmesra.ac.in)

ABSTRACT Brain magnetic resonance images (MRI) convey vital information for making diagnostic decisions and are widely used to detect brain tumors. This research proposes a self-supervised pre-training method based on feature representation learning through contrastive loss applied to unlabeled data. Self-supervised learning aims to understand vital features using the raw input, which is helpful since labeled data is scarce and expensive. For the contrastive loss-based pre-training, data augmentation is applied to the dataset, and positive and negative instance pairs are fed into a deep learning model for feature learning. Subsequently, the features are passed through a neural network model to maximize similarity and contrastive learning of the instances. This pre-trained model serves as an encoder for supervised training and then the classification of MRI images. Our results show that self-supervised pre-training with contrastive loss performs better than random or ImageNet initialization. We also show that contrastive learning performs better when the diversity of images in the pre-training dataset is more. We have taken three differently sized ResNet models as the base models. Further, experiments were also conducted to study the effect of changing the augmentation types for generating positive and negative samples for self-supervised training.

INDEX TERMS Contrastive learning, convolutional neural networks, pre-training, ResNet, self-supervised.

I. INTRODUCTION

The Brain is a complex part of the human body, and any abnormality can affect an individual's health [1]. A brain tumor is an abnormal and uncontrolled growth of the human brain cell. The brain tumor is classified as benign or malignant; or as pituitary, meningioma, or glioma [2], [3]. Invasive approaches such as biopsy or noninvasive methods such as magnetic resonance imaging (MRI), positron emission tomography, and computed tomography are used for detecting brain tumors. Among these, MRI is the most preferred technique due to its capturing detailed information about the tumor's location, progression, shape, and size. To assist a doctor's diagnostic decisions, several researchers have proposed computer-aided systems using machine learning and deep learning methods [4], [5], [6]. Further, several researchers have applied deep learning methods to solve complex problems such as authors in [7] use interpretable deep neural network architecture on sequence-based encoding features for discriminating the Adaptor Protein complexes. Kha et al. [8] proposed a novel model constructed using convolutional neural network (CNN) and position-specific scoring matrix (PSSM) profiles for identification of SNARE proteins.

Deep learning methods such as convolutional neural networks (CNN) do not need manually handcrafted features. They have shown exemplary performance in computer vision on large, labeled datasets such as ImageNet [9]. Such deep models may not be suitable for the medical imaging field, where the sample size of the dataset is usually small. Several researchers have used pre-trained CNN models to overcome this issue and adopted transfer learning and fine-tuning approaches [10], [11], [12], [13]. However, all these approaches apply supervised classification and require a labeled dataset, in the absence of which several researchers have used unsupervised or self-supervised learning. Representation learning through contrastive learning is one such approach; the main idea behind the approach is to learn the representation function by creating augmentations for each

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

data point. Then contrastive loss is applied to maximize the similarity between the data point and its augmentation. At the same time, the similarity between the data point and other samples is minimized. For example, it has been shown [14] that distributions of augmentations of different dog images tend to be similar. Still, their union has little overlap with distributions of augmentations of cat images.

In this work, we apply self-supervised learning based on contrastive loss for brain MRI classification using unlabeled data for pre-training the model. The base architecture used in our experiments is ResNet [15]. Deep networks naturally capture low- and high-level features [16] and classifiers in an end-to-end multilayer fashion, which is further enhanced by increasing the number of layers. However, deep networks come with the problem of vanishing/ exploding gradients, and several researchers have addressed these problems by normalized initialization or intermediate normalization layers [17], [18], [19]. But the degradation problem comes up when the deep networks start converging, and accuracy gets saturated and degrades rapidly [20]. He et al. [15] present a residual learning framework to ease the training of deep networks and address the degradation problem. These ResNet architectures form the base models for our work.

We propose SSCLNet: A **S**elf **S**upervised **C**ontrastive **L**oss based pre-trained **Net**work for Brain MRI classification. The augmented data points are input as positive and negative instances to a deep neural network for label feature learning. The learned features are further passed through a neural network for contrastive learning of instances. Supervised training is applied with a small percentage of labeled data. Finally, the classification is performed using the learned features. We performed numerous experiments with different ResNet architectures and varied the ratio of labeled data used for supervised training. The proposed technique was applied to brain MRI datasets. The major contributions of this paper are as follows: (i) Pre-training a model by self-supervised contrastive learning for Brain MRI classification, and (ii) Performance analysis by varying the percentage of labeled data used for supervised training and changing the augmentation types. The privilege of our work over existing approaches is that the network can learn better features for downstream classification tasks by pre-training. And thus, the SSCLNet proposed by us shows comparable performance to other methods.

The rest of the paper is organized as follows: Section II presents the related work, and Section III presents the methods. Section IV explains the datasets, the implementation details, and the evaluation metrics. Section V presents the results, and Section VI concludes the paper. Our results show that self-supervised pre-training with unlabeled Brain MRI scans improves task performance.

II. RELATED WORK

Computer-aided diagnostic systems have long sought unsupervised learning since labeled data is scarce and expensive, especially in medical image analysis. Over the last decade, deep unsupervised feature learning has been explored to learn the informative representations of images. Most deep unsupervised learning methods aim to learn the feature representations that can reconstruct the inputs themselves, such as the auto-encoder (AE), the sparse auto-encoder (SAE) [21], the denoising auto-encoder (DAE) [22], and the deconvolutional network (DeCNN) [23]. Mishra et al. [24] have applied a semi-supervised approach for generating pseudo labels for classification. Further, deep generative models, including the auto-encoding variational Bayes (AEVB) [25] and the generative adversarial network (GAN) [26], have been provided to encode visual information. Generative adversarial networks have also been utilized for tissue and cell-level categorization, while sparse and variational autoencoders have been employed for unsupervised nuclei detection and transfer learning [27], [28], [29], [30].

Nevertheless, generative models primarily work in the pixel space, which is not scalable. On the other hand, contrastive discriminative methods operate on augmentations of the data point and hence are less expensive computationally. Modern successes in computer vision challenges have lately been attained by contrastive methods based on learning latent-space features by differentiating between unlabeled training data. Such contrastive learning techniques presuppose that two views of the same picture should have comparable feature representations when subjected to minor modifications [31], [32]. The consistency assumption has been exploited by Dosovitskiy et al. to obtain a parametric feature representation for each training instance [33]. Later, Wu et al. [34] extended this work into a non-parametric feature representation using a dynamic memory bank to store latent features of data samples. Any image that is not an augmentation of the original training instance is deemed negative, and the memory bank is utilized to choose negative instances for each training instance. Then, without having to recompute feature vectors, negative samples are obtained using the memory bank. By optimizing the reciprocal information between latent representations of positives, simple picture augmentations (such as resizing images, horizontal flips, color jittering, etc.) and memory banks have successfully learned representations [35], [36]. Ciga et al. [37] have applied self-supervised contrastive loss for digital histopathology datasets. Bootstrap Your Own Latent (BYOL), a novel method for self-supervised image representation learning, is proposed by Grill et al. [38]. It is based on two neural networks, the online and target networks, that communicate with and learn from one another. Contrastive learning's fundamental premise is to transform the original data into a feature space where positive pair similarities are maximized and those of negative pairs are decreased, respectively [39]. The positive and negative pairings are referred to as previous in early writings. Large numbers of data pairs are essential to the effectiveness of contrastive models, as demonstrated by numerous studies [40]. For contrastive learning, several loss functions have been put forward. The distance between an anchor and a positive is minimized while the

distance between an anchor and a negative is increased, for instance, in the case of triplet loss [41]. Nonlinear logistic regression is used in Noise Contrastive Estimation [42] to distinguish between the observed data and some produced noise. SimCLR, a contrastive learning strategy proposed by Chen et al. [43], [44], depends on a large number of mini-batch instances to obtain negative samples for each training instance rather than a custom network or memory bank. Consequently, by supplying more negative samples per training instance over training epochs, the quality of learnt representations was improved.

III. METHODS

The proposed work is based upon the SimCLR approach of Chen et al. [43] and applies contrastive learning of instances for pre-training of the network. Data augmentation operators are applied to data points, then a base encoder learns representations, which are fed into a neural network that maps these representations to a feature space by maximizing agreement between positive examples, as illustrated in Figure 1. The SSCLNet architecture is split into three blocks – the Label Feature Generation (LFG) Block, the Instance Level Contrastive Learning (ILCL) Block, and finally, the Supervised Classification (SC) Block.

A. LABEL FEATURE GENERATION (LFG) BLOCK

The proposed framework uses data augmentation to construct data pairs. Given a data instance d_i, two transformations Γ^a *and* Γ^b are applied, resulting in $d_i^a = \Gamma^a(d_i)$ and $d_i^b = \Gamma^b(d_i)$. In our work, the data augmentations used are as follows: random cropping, random brightness, random contrast, and random noise. One shared deep neural network σ (.) is used to extract label features from the augmented samples as follows: $l_i^a = \sigma(d_i^a)$ and $l_i^b = \sigma(d_i^b)$. In our work, three Resnet architectures have been used; however, the method does not depend on any specific network.

B. INSTANCE LEVEL CONTRASTIVE LEARNING (ILCL) BLOCK

Contrastive learning aims to maximize the similarities of positive pairs while minimizing them for negative pairs. Positive pairs in our work are defined as those generated from the same instance, and negative pairs otherwise. Thus for a mini-batch of size M, two types of augmentations are performed on each instance d_i, and $2M$ data samples are generated as, $\left\{d_1^a, d_2^a, \ldots, d_M^a, d_1^b, d_2^b, \ldots, d_M^b\right\}$. For a specific sample d_i^a, there is one positive pair $\left\{d_i^a, d_i^b\right\}$, and the remaining $2M - 2$ are negative pairs.

In this block, for contrastive instance level learning, we take a four-layer nonlinear multilayer perceptron α (.) to map the features l_i learnt from the LFG Block to a subspace $z_i^a = \alpha(l_i^a)$ and $z_i^b = \alpha(l_i^b)$ where the instance level contrastive loss is applied. The pairwise similarity is measured as $s\left(z_i^{k_1}, z_j^{k_2}\right) = z_i^{k_i}.z_j^{k_2}$ where $r.s$ denotes the *dot product of* r *and* s; $k_1, k_2 \in \{a, b\}$; and $i, j \in [1, M]$. The loss e_i^a for a given sample d_i^a is given as,

$$e_i^a = -log\frac{\left[exp\left(s(z_i^a, z_i^b)/\tau_1\right)\right]}{\sum_{j=1}^{M}\left[\exp\left(s(z_i^a, z_j^a)/\tau_1\right) + \exp((z_i^a, z_j^b)/\tau_1)\right]} \tag{1}$$

where τ_1 is the *temperature parameter*.

Additionally, we put a constraint on the derived features, such that, the $L_2 - norm$ of the vector is 1. That is, $\forall i$, $\|l_i\|_2 = 1$, and $l_{iz} \geq 0$, $z = 1, \ldots, y$, where, $\|.\|_2$ represents the $L_2 - norm$ of a vector and l_{iz} is the z^{th} element of label feature l_i.

The instance level contrastive loss L_i is calculated for every augmented sample as,

$$L_i = \frac{1}{2M}\sum_{i=1}^{M}(e_i^a + e_i^b). \tag{2}$$

C. CLASSIFICATION (CL) BLOCK

The features learned from the LFG and the ILCL blocks are applied for classification in the Classification Block, which comprises a neural network $\emptyset$. The loss function used is the categorical cross-entropy loss.

D. DESCRIPTION OF SSCLNet

The essential feature of the proposed approach is the learning of representations by means of positive and negative samples. Given an image x, augmentations are applied to it to generate samples x' and x''. Now, the image pairs $\left(x, x'\right)$, $\left(x', x''\right)$ and $\left(x, x''\right)$ are treated as positive samples. For all other images $y \neq x$, the pairs (x, y) are treated as negative samples. This has also been presented in Figure 2. The contrastive loss function maximizes the agreement between positive samples while minimizing the agreement between negative samples. This concept has been implemented by the SSCLNet architecture proposed in this work. In the Label Feature Generation (LFG) Block, embeddings are generated for the augmented data pairs by the shared network σ (.) which comprises of the ResNet architecture. In the Instance Level Contrastive Learning (ILCL) Block, a four layer multilayer perceptron α (.) is used and contrastive loss is applied. The pairwise similarity of positive samples is increased while that of negative pairs is reduced. These learnt features are then input to the final classification layer $\emptyset$ (.) with categorical cross entropy loss, for obtaining the output.

IV. EXPERIMENTS

A. DATASETS

Two datasets from the Kaggle repository [45], [46] have been used in this study. However, we created our own datasets by applying augmentation to the Brain MRI 2-Class and 4-Class datasets from the Kaggle repository. The dataset of Brain MRI Tumor 2-Class used in this study has 2580 normal samples and 2561 tumor samples in the training dataset and 651 normal samples and 634 tumor samples in the test set. In this study, the images were made grayscale, and the border of the skull was located by erasing the background color from

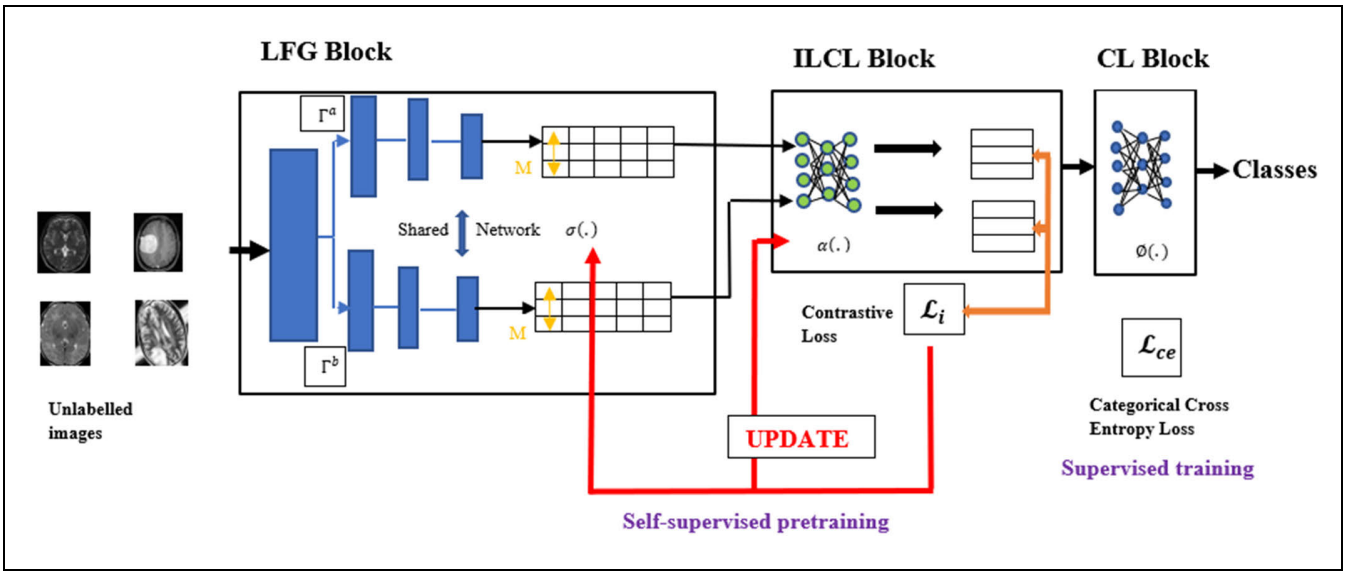

FIGURE 1. The block diagram of the proposed framework. The images are subject to two data augmentations, and features are learned by shared networks in the Label Feature Generation Block. A neural network in the Instance Level Contrastive Learning Block projects the features for maximizing agreement by contrastive loss. The features from this embedding network are fed into the Classification Block for classification.

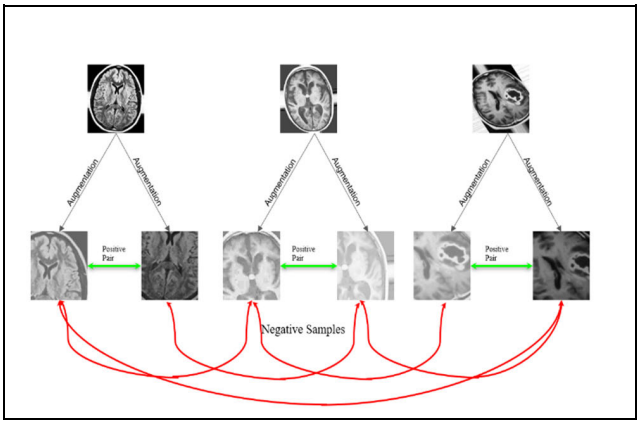

FIGURE 2. Positive and negative samples.

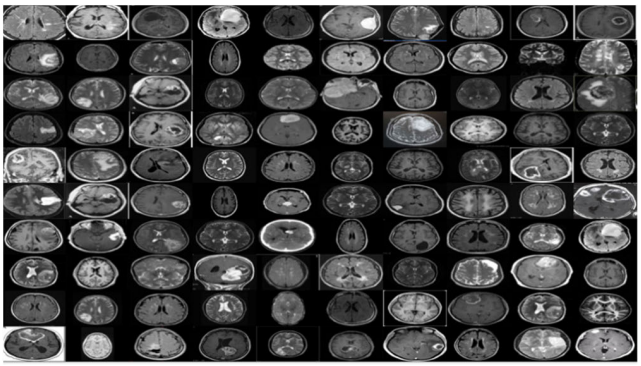

FIGURE 3. Brain MRI 2-class dataset visualization.

the image. As a result, it offered the original image's contour. Histogram equalization and median filters were used. The original dataset contains 512 × 512 images in various dimensions. All of these were downsized to 224 × 224 for processing and normalized between 0 and 1. The dataset of Brain MRI Tumor 4-Class used in this study contains 826 samples of glioma tumor, 822 samples of meningioma tumor, 395 samples of no tumor, and 827 samples of pituitary tumor class in the training set, and 100 samples of glioma tumors, 115 samples of meningioma tumors, 105 samples of no tumors, and 74 samples of pituitary tumors for testing purposes. Median filters and histogram equalization were applied. The original size of the dataset was 512×512, which we resized into 224×224.

The visualization of 2-class and 4-class datasets are given in Figures 3 and 4, while Figures 5 and 6 present the augmented images. The augmentations have been chosen randomly from the following: random cropping, random brightness, random contrast, and random noise.

B. IMPLEMENTATION DETAILS

We adopted several ResNet architectures (18, 34, and 50) as our backbone architecture. For each architecture, three sets of experiments were conducted. The first one with random initialization of the ResNet, the second one with ImageNet initialization, and the third one was the SSCLNet. For this, for initial adapting, we fine-tuned our ResNet architectures for 100 epochs by adding a few layers after convolutional layers. The hyperparameter tuning was done by running several experiments. Our next step was the contrastive pre-training step. For the contrastive learning framework, we used four dense layers with 512, 512, 256, and 256 neurons, with a dropout of 0.4 in the last dense layer. The network comprising ResNet architecture and the dense layers is named the

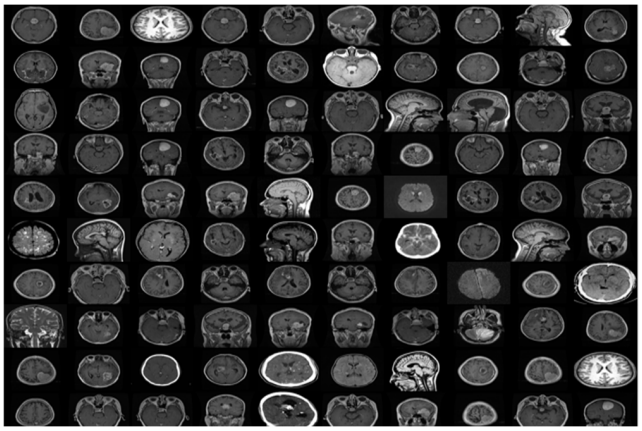

FIGURE 4. Brain MRI 4-class dataset visualization.

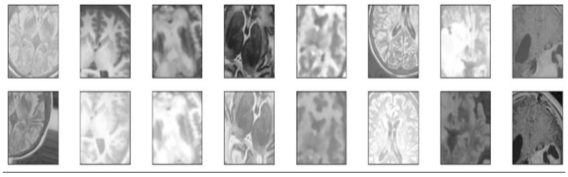

FIGURE 5. Augmented images of Brain MRI 2-class dataset.

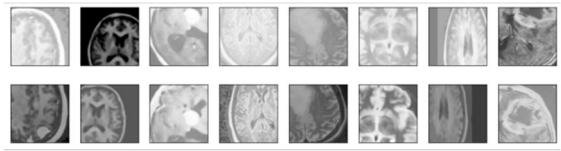

FIGURE 6. Augmented images of Brain MRI 4-class dataset.

Embedding network. The output dimension of the embedding layers was fixed at 32. The entire embedding network is then trained end to end in a self-supervised fashion. Adam optimizer with an initial learning rate of 0.0003 was adopted. Owing to memory limitations, we fixed the batch size at 64. The next phase is the supervised training phase. Here we initialize our supervised architecture consisting of dense layers with SoftMax as the output layers. Seven dense layers with dimensions 256, 256, 128, 128, 64, 32, and 16 were applied before the final classification SoftMax layer. The representations generated from the previous embedding network act as input to the supervised architecture. The SoftMax prediction of the vector representation acts as the output. With the labels of the embeddings, the supervised architecture is then trained end to end as supervised training. The amount of labeled data available was varied to check the model's performance for different percentages of labeled data. We compare the pre-trained proposed network SSCLNet with randomly initialized and ImageNet pre-trained ResNet 18, 34, and 50. We have adopted fine-tuning for supervised training. The widely used metrics accuracy, F1-score, precision, and recall are used to evaluate our method. Higher values of these metrics indicate better performance.

For the pre-training dataset, we randomly sample images from the 2-Class and 4-Class datasets. The implementation code can be found at [49].

V. RESULTS AND ANALYSIS

We compare self-supervised pre-trained networks with random and ImageNet initialization for ResNet 18, ResNet 34, and ResNet 50.

A. OVERALL PERFORMANCE ANALYSIS

It is seen from the graph plots of Figure 7 that the self-supervised pre-trained network, SSCLNet is superior to ImageNet initialization for the 4-Class dataset. SSCLNet gives the highest accuracy of 63.45%, 53.3%, and 69.04% for the ResNet 18, ResNet 34, and ResNet 50 architectures. The F1-Scores for SSCLNet are 68%, 56%, and 75% for the three architectures, achieving the highest value in all three cases. The results of SSCLNet applied to the 2-Class dataset (presented in Figure 8) show the highest values of accuracy and F1-Scores for the ResNet 50 architecture and not-so-promising values for the others. The ROC curves presented in Figure 9 also show that the SSCLNet architecture gives the best AUC value for the Brain MRI 2-class data for ResNet 50 model.

B. PERCENTAGE LABELED DATA FOR SUPERVISED TRAINING

We conducted experiments with ResNet 50 architecture to study the variation in accuracy and F1-score performance when the ratio of labeled data used for supervised training is changed. For the 4-Class dataset, we find that at 30% labeled data, the accuracy values are 43% for Random, 45% for ImageNet initialization, and 48% for SSCLNet. At 50% labeled data, the values for Random initialization and SSCLNet show an increase, 52% and 51%, respectively, but that for ImageNet initialization shows a fall from 45% to 44%. Though the accuracy curve for the Random and ImageNet initialization shows a zig-zag pattern indicating a fall in accuracy for an increased percentage of labeled data, the curve for SSCLNet shows a constant upward movement, as shown in Figure 10. Similar behavior is observed in the F1-Score curve. Thus, we can say that there is an increase in the performance of SSCLNet with the increase in the percentage of labeled data used for supervised training, and the increase is from 48% accuracy at 30% labeled data to 69% accuracy for 100% labeled data. However, for the 2-Class dataset, as shown in Figure 11, the SSCLNet does not have a very smooth upward curve, even though there is an overall increase in accuracy from 63% at 30% data to 71% at 100% labeled data. A similar increase from 63% at 30% data to 71% at 100% labeled data is seen for the F1-Score as well.

C. EFFECT OF AUGMENTATION TECHNIQUES

We experimented by applying different augmentation techniques with the ResNet 50 backbone for SSCLNet, and the 4-Class dataset.

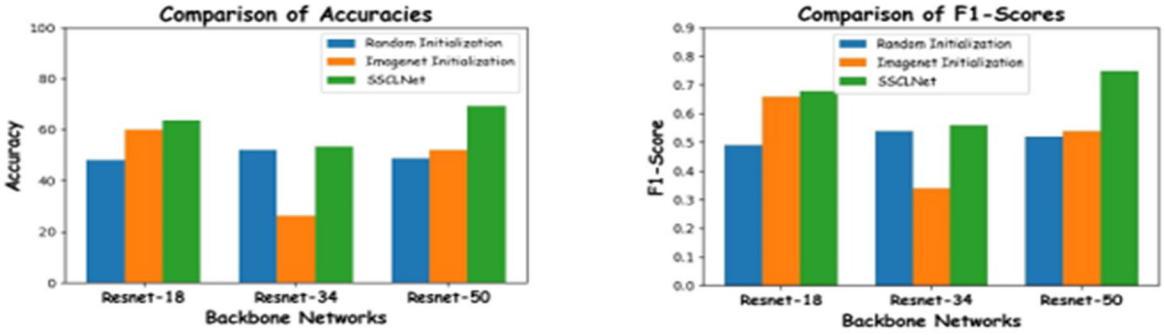

FIGURE 7. Overall performance analysis for Brain MRI 4-class dataset.

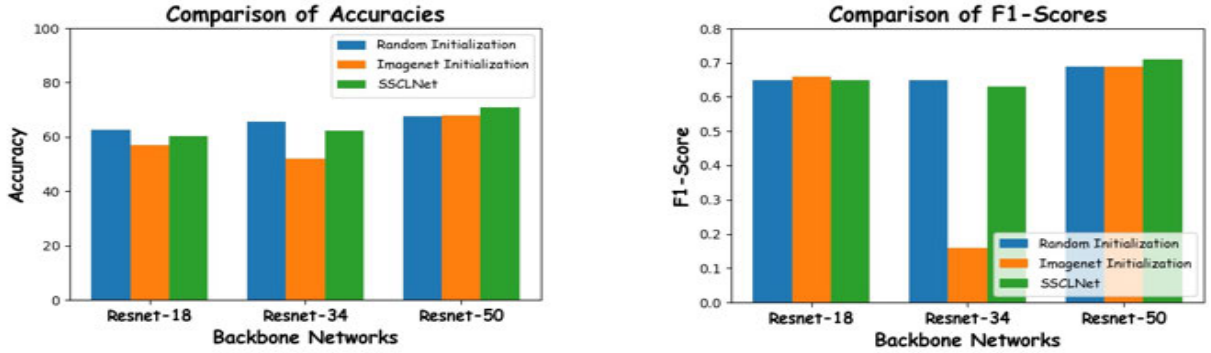

FIGURE 8. Overall performance analysis for Brain MRI 2-class dataset.

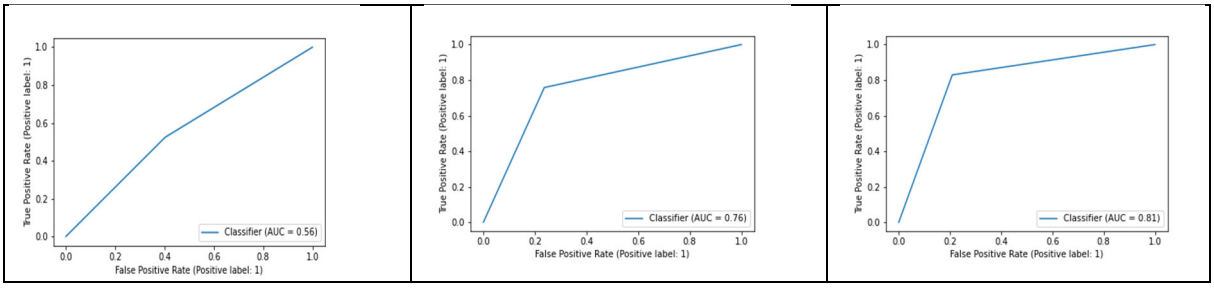

FIGURE 9. ROC curves for Random, Imagenet and SSCLNet initialization for Brain MRI 2-class dataset, ResNet 50 architecture.

TABLE 1. Effect of augmentation techniques.

Augmentations	Accuracy (%)	F1- Score (%)
Cropping	68.27	72
Cropping, contrast	66.5	69
Cropping, brightness	69.8	73
Cropping, Noise	64.97	70
brightness, noise	69.54	72
Cropping, brightness, noise, contrast	69.04	75

The results of experiments with different augmentation techniques are presented in Table 1, and it is seen that by selecting randomly from different augmentation techniques, the accuracy and F1-Score values vary from 64.97% to 69.04% and from 69% to 75%, respectively. This shows that the choice of augmentation techniques has an influence on the performance of the model and may be done based on a validation subset. In the present study, random cropping, random brightness, random contrast, and random noise have been applied, and the results are shown in the last row of Table 1. The F1-Score is the highest among all the experiments; however, accuracy is marginally low (by 0.5%) from the random brightness and random noise techniques.

D. STATISTICAL EVALUATION

An interval statistic called a confidence interval (CI) is used to express how uncertain an estimate is. It offers both, a like-

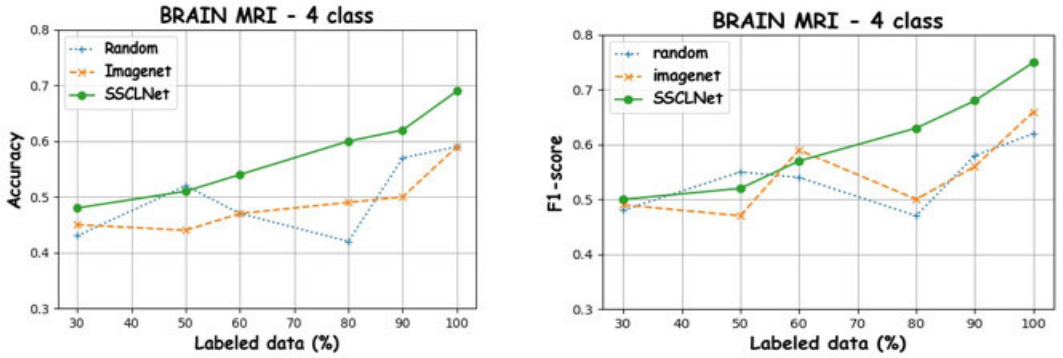

FIGURE 10. Accuracy and F1 score plots when {30, 50, 60, 80, 90, 100} % of the labeled Brain MRI – 4 class data is used for supervised training.

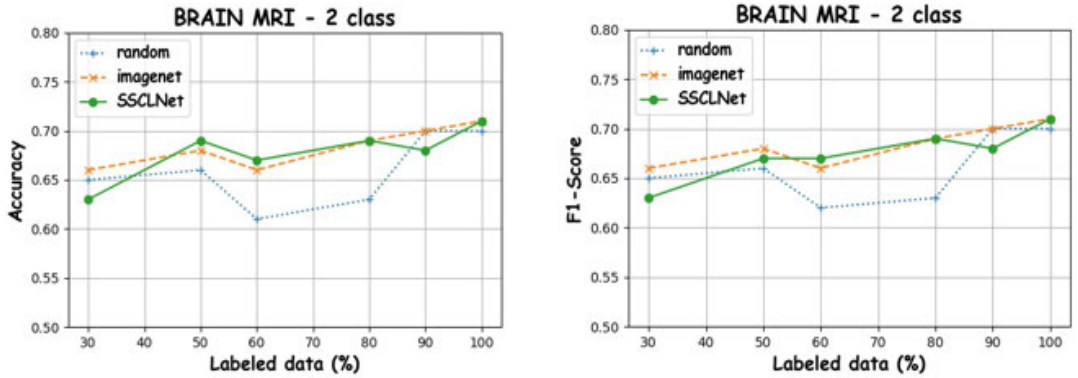

FIGURE 11. Accuracy and F1 score plots when {30, 50, 60, 80, 90, 100} % of the labeled Brain MRI – 2 class data is used for supervised training.

TABLE 2. 95% CI for ResNet 18 architecture, 4-class.

Initialization	Accuracy	CI
Random	48%	±0.30%
Imagenet	60%	±0.20%
SSCLNet	63.45%	±0.20%

TABLE 3. 95% CI for ResNet 34 architecture, 4-class.

Initialization	Accuracy	CI
Random	52.03%	±0.09%
Imagenet	26.14%	±0.2%
SSCLNet	53.3%	±0.09%

lihood, a lower and upper bound. According to Cumming & Calin-Jageman [47], a short CI typically denotes a tiny margin of error. This range can be used to calculate a model's capability estimate. In addition to statistical significance tests, CI is a branch of statistics that can be used to report and evaluate experimental results [48]. The typical calculations put them at 95%, 98%, and 99%. According to a 95% CI, 95% of the studies conducted will fall inside the range, whereas 5% will not. We now give the calculated results of the numerous experiments presented earlier, with the Brain MRI 4-class dataset, with a 95% confidence interval in Table 2 – Table 4.

TABLE 4. 95% CI for ResNet 50 architecture, 4-class.

Initialization	Accuracy	CI
Random	48.73%	±0.09%
Imagenet	52.03%	±0.1%
SSCLNet	69.04%	±0.30%

It is seen from Table 2 that for the ResNet 18 architecture and 4-class dataset, the accuracy values for SSCLNet architecture vary from 63.314% to 63.586% that is, $(63.45 \pm 0.136)\%$. This implies that it is expected that with 95% confidence the efficiency of the proposed model is likely between 63.314% and 63.586%.

From Table 3, it is seen that for ResNet 34 architecture and the 4-class dataset, the accuracy values for SSCLNet architecture vary from 53.252% to 53.348% that is, $(53.3 \pm 0.0482)\%$. This implies that it is expected that with 95% confidence the efficiency of the proposed model is likely between 53.252% and 53.348%.

From Table 3, it is seen that for ResNet 50 architecture and the 4-class dataset, the accuracy values for SSCLNet architecture vary from 68.867% to 69.213% that is, $(69.04 \pm 0.173)\%$. This implies that it is expected that with 95% confidence the efficiency of the proposed model is likely between 68.867% and 69.213%.

E. DISCUSSION

From the accuracy and F1-Score values of 2-Class and 4-Class data presented in Figures 7 and 8, we see that contrastive learning shows remarkable improvement when the pre-training dataset contains more diverse images, which is the case with the 4-Class dataset (Figure 4). The increase in the percentage of labeled data used for supervised training also enhances the performance of SSCLNet. Similarly, changing the augmentations applied to data samples impacts the accuracy and F1-Scores, as seen in Table 1. The ROC curves presented in Figure 9 also show the better performance of SSCLNet architecture. It is further noted from the results in Figures 10 and 11, that the improvement in accuracy is approximately 10% for the SSCLNet, as compared to the other two initialization methods. In addition, the results of the experiments with the Brain MRI 4-class dataset, with a 95% confidence interval as presented in Table 2 – Table 4 show that our results are very stable, varying between 63.314% and 63.586% for the ResNet 18, between 53.252% and 53.348% for ResNet 34 and between 68.867% and 69.213% for the ResNet 50 architecture. These findings force one to ponder upon the following questions:

- Can one expect to find improvements if the pre-training dataset is made by sampling from both 2-Class and 4-Class data samples?
- Can we use the learned representations for clustering tasks, and will there be an improvement in performance?
- What would be the effect of increasing the size of the pre-training dataset?

These we would like to investigate in our future works.

VI. CONCLUSION

It is important that good features are learned to achieve good performance in complex tasks like computer vision or pattern recognition. In our work, contrastive learning has been applied for learning the instances by which the model is pretrained with unlabeled data, and this is used for the classification of Brain MRI images. To our knowledge, developing a classification model using unlabeled data and self-supervised learning for MRI classification has not been done prior to this work. Our proposed SSCLNet applies the SimCLR approach, which learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. A stochastic data augmentation module transforms any given data example randomly, resulting in two correlated views of the same example, which are treated as the positive pair. Then neural network encoders are applied to extract representations from augmented examples. Supervised training is done using labeled data, and then the model is used for the classification of Brain MRI images. In this work, our aim was to show that by pre-training, better features can be learned for downstream classification tasks. The SSCLNet proposed by us shows comparable performance to ImageNet training. It is also found that contrastive learning may not show much improvement when representations fail to encode domain-specific information due to a smaller number of negative samples or when there is lesser variation in the pre-training dataset.

ACKNOWLEDGMENT

The authors are indebted to, and thank the anonymous reviewers for providing valuable suggestions, which helped them prepare the article in its present form.

REFERENCES

[1] R. Hoshide and R. Jandial, "2016 world health organization classification of central nervous system tumors: An era of molecular biology," *World Neurosurg.*, vol. 94, pp. 561–562, Oct. 2016, doi: 10.1016/j.wneu.2016.07.082.

[2] *American Cancer Society*. Accessed: Jul. 21, 2022. [Online]. Available: https://www.cancer.org/cancer.html

[3] *Brain Tumor: Diagnosis*. Accessed: Jul. 21, 2022. [Online]. Available: https://www.cancer.net/cancer-types/brain-tumor/diagnosis

[4] G. S. Tandel, M. Biswas, O. G. Kakde, A. Tiwari, H. S. Suri, M. Turk, J. R. Laird, C. K. Asare, A. A. Ankrah, N. N. Khanna, B. K. Madhusudhan, L. Saba, and J. S. Suri, "A review on a deep learning perspective in brain cancer classification," *Cancers*, vol. 11, no. 1, p. 111, 2019, doi: 10.3390/cancers11010111.

[5] M. M. Badža and M. Č. Barjaktarović, "Classification of brain tumors from MRI images using a convolutional neural network," *Appl. Sci.*, vol. 10, no. 6, p. 1999, Mar. 2020, doi: 10.3390/app10061999.

[6] W. Anjali, B. Anuj, and V. S. Verma, "A review on brain tumor segmentation of MRI images," *Magn. Reson. Imag.*, vol. 61, pp. 247–259, Sep. 2019, doi: 10.1016/j.mri.2019.05.043.

[7] Q.-H. Kha, T.-O. Tran, T.-T.-D. Nguyen, V.-N. Nguyen, K. Than, and N. Q. K. Le, "An interpretable deep learning model for classifying adaptor protein complexes from sequence information," *Methods*, vol. 207, pp. 90–96, Nov. 2022, doi: 10.1016/j.ymeth.2022.09.007.

[8] Q.-H. Kha, Q.-T. Ho, and N. Q. K. Le, "Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles," *J. Chem. Inf. Model.*, vol. 62, no. 19, pp. 4820–4826, Sep. 2022, doi: 10.1021/acs.jcim.2c01034.

[9] O. Russakovsky, J. Deng, H. Su, and J. Krause, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015, doi: 10.1007/s11263-015-0816-y.

[10] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, and Z. Xu, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[11] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519, doi: 10.1109/CVPRW.2014.131.

[12] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–45, doi: 10.1109/CVPRW.2015.7301270.

[13] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51, doi: 10.1109/CVPRW.2015.7301382.

[14] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1–19.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[16] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[17] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[18] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[20] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5353–5360.
[21] A. Ng, "Sparse autoencoder," *CS294A Lect. Notes*, vol. 72, pp. 1–19, Jan. 2011.
[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.
[23] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.
[24] A. Mishra and V. Bhattacharjee, "Applying semi-supervised learning on human activity recognition data," in *Proc. Int. Conf. IoT Blockchain Technol. (ICIBT)*, May 2022, pp. 1–6, doi: 10.1109/ICIBT52874.2022.9807808.
[25] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[26] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[27] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 119–130, Jan. 2016, doi: 10.1109/TMI.2015.2458702.
[28] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, May 2018, doi: 10.1109/TPAMI.2017.2656884.
[29] L. Hou, V. Nguyen, A.B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, and D. Foran, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images," *Pattern Recognit.*, vol. 86, pp. 188–200, Feb. 2019, doi: 10.1016/j.patcog.2018.09.007.
[30] B. Hu, Y. Tang, E. I.-C. Chang, Y. Fan, M. Lai, and Y. Xu, "Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1316–1328, May 2019, doi: 10.1109/JBHI.2018.2852639.
[31] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2794–2802, doi: 10.1109/ICCV.2015.320.
[32] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, no. 6356, pp. 161–163, Jan. 1992, doi: 10.1038/355161a0.
[33] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016, doi: 10.1109/TPAMI.2015.2496141.
[34] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742, doi: 10.1109/CVPR.2018.00393.
[35] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15509–15519.
[36] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. Ali Eslami, and A. van den Oord, "Data-efficient image recognition with contrastive predictive coding," 2019, *arXiv:1905.09272*.
[37] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," 2020, *arXiv:2011.13971*.
[38] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.
[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.
[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[41] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[42] M. Gutmann and A. Hyvarinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.
[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.
[44] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," 2020, *arXiv:2006.10029*.
[45] *MRI 2 Class Dataset*. Accessed: Apr. 10, 2022. [Online]. Available: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection
[46] *MRI 4 Class Dataset*. Accessed: Apr. 10, 2022. [Online]. Available: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri
[47] G. Cumming and R. Jageman, *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Evanston, IL, USA: Routledge, 2016.
[48] A. Claridge-Chang and P. N. Assam, "Estimation statistics should replace significance testing," *Nature Methods*, vol. 13, no. 2, pp. 108–109, Jan. 2016, doi: 10.1038/nmeth.3729.
[49] A. Mishra. *SSCLNet*. Accessed: Dec. 29, 2022. [Online]. Available: https://github.com/cheersanimesh/SSCLNet

ANIMESH MISHRA is currently pursuing the undergraduate degree with the Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi. He has a passion for coding. He is highly interested in various research areas of computer science and would like to pursue higher studies. His current interest includes explore the emerging areas in machine learning.

RITESH JHA received the M.Sc. degree in computer science from G. B. Pant University and the Ph.D. degree in computer science from BIT, Mesra, Ranchi, India. Currently, he is an Assistant Professor with the Department of Computer Science and Engineering, BIT. His current research interest includes machine learning applied to healthcare data.

VANDANA BHATTACHARJEE received the B.E. degree in CSE from the Birla Institute of Technology (BIT), Mesra, Ranchi, in 1989, and the M.Tech. and Ph.D. degrees in computer science from Jawaharlal Nehru University, New Delhi, in 1991 and 1995, respectively. She is a Professor with the Department of Computer Science and Engineering, BIT. She has several national and international publications in journal and conference proceedings. She has coauthored a book on data analysis. Currently, she is working on deep learning techniques applied to the domains of software fault prediction, classification of images, disease prediction, and learning without labels. Her research interests include machine learning and its applications. She is a Life Member of Computer Society of India.