Improving Brain Tumor Classification Using Semi-Supervised Latent Space Learning

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

In

Computer Science and Engineering

Submitted By

SARANYA BHATTACHARJEE UNIVERSITY ROLL NO: 12019009001311

NISHA KUMARI UNIVERSITY ROLL NO: 12019009023082

ANIKET SINGH UNIVERSITY ROLL NO: 12019009022126

Under the guidance of

Prof. Sankhadeep Chatterjee &

Prof. Amartya Chakraborty

Department of Computer Science and Engineering

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

UNIVERSITY
OF ENGINEERING & MANAGEMENT
'University Area', Plot No. III-B/5, Main Arterial Road, New Town, Action Area-III, Kolkata - 700 160, W.B., India
City Office : 'ASHRAM', GN-34/2, Salt Lake Electronics Complex, Kolkata - 700 091, W.B., India
(Established by Act XXV of 2014 of Govt. of West Bengal & recognised by UGC, Ministry of HRD, Govt. of India)

Ph. (Office) : 91 33 2357 7649
: 91 33 2357 2969
: 91 33 6888 8606
Admissions : 91 33 2357 2059
Fax : 91 33 2357 8302
E-mail : vc@uem.edu.in
Website : www.uem.edu.in

CERTIFICATE

This is to certify that the project titled "**Improving Brain Tumor Classification using Semi-Supervised Latent Space Learning**" submitted by, **SARANYA BHATTACHARJEE** (University Roll No: 12019009001311), **NISHA KUMARI** (University Roll No: 12019009023082), and **ANIKET SINGH** (University Roll No: 12019009022126)**,** students of the UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfillment of the requirement for the degree of Bachelor of Computer Science Batch 2023, is a bonafide work carried out by them under the supervision and guidance of Prof. Sankhadeep Chatterjee and Prof. Amartya Chakraborty during 8th Semester of the academic session of 2023. The content of this report has not been submitted to any other university or institute. I am glad to inform you that thework is entirely original and its performance is found to be quite satisfactory.

Sankhadeep Chatterjee. 10/5/23

Prof. Sankhadeep Chatterjee
Department of Computer Science & Technology
UEM, Kolkata

Amartya Chakraborty

Prof. Amartya Chakraborty
Department of Computer Science & Technology
UEM, Kolkata

Prof. Sukalyan Goswami
HOD, Department of Computer Science
UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remain invaluable to us.

We are sincerely grateful to my guides Prof. Sankhadeep Chatterjee and Prof. Amartya Chakraborty of the Department of Computer Science, UEM, Kolkata, for their wisdom, guidance, and inspiration helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof. Sukalyan Goswami, UEM, Kolkata, and all other departmental faculties for their ever-present assistance and encouragement.

Last but not least, we would like to extend our warm regards to my family and peers who have kept supporting me and always had faith in our work.

- Saranya Bhattacharjee
- Nisha Kumari
- Aniket Singh

TABLE OF CONTENTS

ABSTRACT

Recent research has emphasized the significance of classifying brain tumors to enable timely diagnosis and treatment. Magnetic Resonance Imaging (MRI), a non-invasive imaging technique, has been extensively utilized to capture the growth of tumors in brain tissue. However, limited labeled data poses a significant challenge in training, effective deep learning models. This study addresses this issue by utilizing a semi-supervised latent space learning approach that leverages a large amount of unlabeled data. Initially, a Variational Autoencoder (VAE) is used to learn the optimal latent vector representation of brain MR images. Then, a shallow learner is trained using a small set of labeled latent vectors. Later, a pool of unlabeled latent vectors, obtained from a previously trained VAE encoder, is employed to fine-tune the partially trained model using a self-training strategy of semi-supervised learning. The proposed framework has been extensively evaluated in terms of Accuracy, Precision, Recall, and F1-Score using two distinct datasets for binary and multi-class tumor type classification. To achieve the most promising results in our model, separate experiments were conducted to determine the optimal initial size of the labeled set and the best-performing base classifier. Moreover, the performance of the proposed model was compared with fully-supervised baselines and state-of-the-art semi-supervised brain tumor detection frameworks.

SECTION – I: INTRODUCTION

Over the last few decades, automated brain tumor classification has become one of the most explored subjects in clinical examination [1], [2]. Research indicates that early detection of brain tumors, malignant (cancerous) or benign, can spare the patient from serious complications. Non-invasive techniques like computed tomography (CT) and magnetic resonance imaging (MRI) have been the two most common modalities exploited to identify aberrant tissue growth in the brain. Typically, in traditional practices, radiologists utilize manual diagnosis to detect tumor growth. Such brain scan assessments are frequently time-consuming and expensive procedures, which can occasionally result in misclassification. Thus, to enhance the diagnostic ability and accelerate the process of detecting, segmenting, and identifying the type of brain tumor, computer-aided diagnosis (CAD) systems have been immensely valuable [3]. In recent years, Deep Learning (DL) methodologies have been extensively studied in medical imaging [4]–[6]. Especially, in supervised learning scenarios where there is often a requirement for a huge amount of data, DL algorithms have had significant breakthroughs [7], [8]. However, acquiring high-quality annotated data samples is difficult and demands expert supervision. With few labeled instances, it becomes challenging to develop a reliable fully supervised system. Unlabeled samples, on the contrary, are abundant and easily available. Consequently, leveraging the value of these unlabeled datasets and annotating them without manual intervention is a promising solution. For this purpose, Semi-supervised learning (SSL) has recently sparked a lot of interest from the research community due to its ability to effectively boost overall performance and reduce human annotation costs to a great extent [9].

SSL is a learning paradigm involving the use of unlabeled and labeled data. Concerning medical image analysis, there have been several studies employing SSL in the last few years [10]. In one such study, Ge et al. [11] proposed a novel 3D-2D consistent graph-based deep semi-supervised learning framework for glioma-type classification. Meanwhile, generative models like Variational Autoencoders (VAEs) [12] and Generative Adversarial Networks (GANs) [13] have exhibited the strength of interpretable deep features which can not only enhance the performance of a model but can provide insightful hidden information about the data [14]– [18]. As a result, in this present article, we have exploited the representation and statistical learning ability of VAE to obtain the most effective underlying feature representations of brain MR images and implemented a self-trained SSL framework to boost the performance of a base classifier trained on a limited amount of labeled data.

SECTION – II: LITERATURE SURVEY

Recent years have demonstrated the superiority of deep generative models. Among them, Variational Autoencoders (VAEs) [12] have gathered significant popularity as a low-dimensional manifold representation learning approach. A significant advantage of VAEs is the power to control the distribution of the latent representation, which tends to become very beneficial in downstream tasks [19], [20]. Primarily, having an encoder-decoder Bayesian network architecture, VAEs can learn smooth latent representations of the input data. In medical image processing, too, VAEs have been well-researched [21]–[24]. In [25], the authors extracted both the radiomic and the shape features learned by a VAE, to detect the MGMT Methylation Status of Glioblastoma. In another study [26], the authors employed a convolutional VAE (CVAE) to convert a small class imbalanced dataset to a large balanced one for brain tumor detection and classification. The authors in [27] proposed an adversarial training framework consisting of a generator, a discriminator, a latent regularize, and an auxiliary encoder to study different types of brain lesions.

Several SSL methods have been proposed in recent years that leverage the huge amount of unlabeled data to improve the performance of machine learning models [28], [29]. Some of its applications can be seen in domains, such as agriculture [30], [31], medical research [10], [32], emotion recognition [33], [34], and text classification [35]. Among the various methods, the two of the most well-known inductive SSL algorithms are self-training [36] and co-training [37]. Unlike self-training, where the classifier teaches itself by exploiting its predictions, co-training attempts to use the mutual information between two individual models trained on different views of the data. In [38], the authors proposed an ensemble of auto-encoding transformations using the self-training approach. In another study, a self-trained SSL methodology based on knowledge distillation was put forward [39]. For medical image classification, the authors in [40] presented a novel deep virtual adversarial self-training SSL framework incorporating consistency regularization. VAE-based SSL frameworks are also very effective [41]. Another semi-supervised strategy by [42] calculated the dissimilarities between latent representations obtained by a VAE to classify brain MRIs as healthy and unhealthy. In [43], the authors proposed a novel framework incorporating binary angular learning and a semi-supervised binary classifier. Extensive evaluation of the framework on a binary brain tumor dataset demonstrated the effectiveness of a semi-supervised classifier in obtaining promising accuracy compared with a fully supervised setting. The authors in [44] proposed a novel semi-supervised hierarchical multitask learning framework based on multi-modal brain MRI. In another study [45], two

improved semi-supervised Expectation filtering maximization and MCo_Training classifiers were proposed. Using these classifiers and a graph-based semi-supervised classifier as components in an ensemble framework, the authors demonstrated noteworthy results in terms of accuracy and precision. In [46], the authors explored semi-supervised Support Vector Machines for brain image fusion. The proposed framework improved recognition performances while securing low error rates.

SECTION – III: MOTIVATION & OBJECTIVE

The literature survey has revealed that brain tumor segmentation and detection of specific diseases may be a difficult task due to the absence of good quality labeled data required in the training of deep learning models. Motivated by this, in the current study, the authors have proposed a self-learning semi-supervised learning framework to efficiently detect glioma and meningioma diseases. However, to improve the performance of the proposed semi-supervised learning method, a variational autoencoder is utilized to learn the most effective latent vector representation of input MR images. The latent vectors (both labeled and unlabeled) are then used to train the semi-supervised learning model. The base classifier of the model is varied to find out the most suitable base classifier for detecting brain tumors. Overall, the contributions of the current manuscript are as follows:

1) Variational Autoencoder is trained to obtain the best latent representation of input brain MR image.
2) The latent vectors of both labeled and unlabeled images are used to train the semi-supervised model.
3) The base classifier is varied and a set of four distinct classifiers are tested in terms of Accuracy, Precision, Recall, and F1-Score.
4) Separate experiments are conducted to determine the optimal initial size of the labeled set.
5) Comparison with fully-supervised baselines, state-of-the-art methods, and statistical significance tests are performed to establish the effectiveness of the proposed methodology.

SECTION – IV: METHODOLOGY

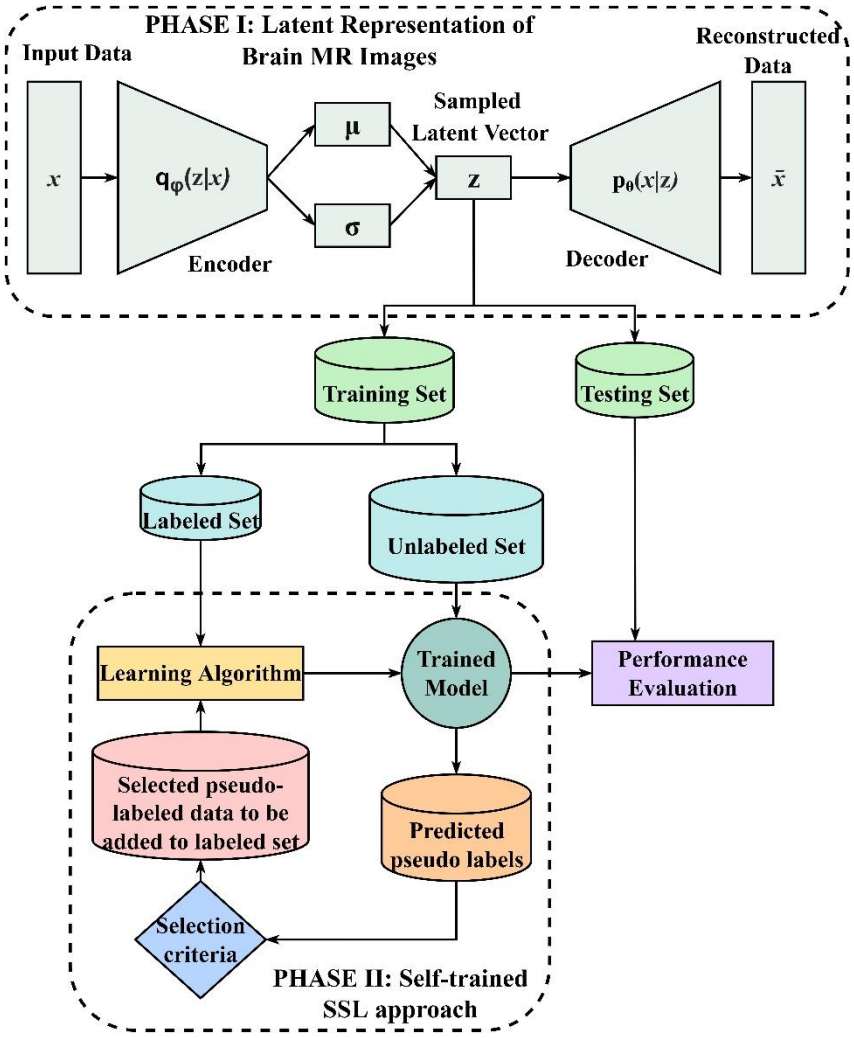

FIGURE 1: Flow Diagram

A. LATENT REPRESENTATION OF BRAIN MR IMAGES

A great deal of research in generative modeling, conducted in the last few years, has elaborated on how the most important features in latent space can be useful for improving classification performance [16], [17], [47], [48]. The intuition behind this is that if the VAE is properly trained to reconstruct the data, then the latent vector retains the meaningful and critical information of the original data. Motivated by this, the first step of our proposed architecture, as demonstrated in Fig. 1, is to employ a VAE architecture to acquire compact and low-dimensional latent representations of the brain MR images. The encoder network, as illustrated in Fig. 2, has 2 convolutional layers with ReLU activation function. A max-pooling layer follows each convolutional layer. The last layer

of our architecture is a dense layer that encodes twice the size of the latent space. Further, the VAE maps the input data points to a multivariate normal distribution, which is parameterized by mean μ and log-variance σ vectors (both having the same dimensionality of the latent space). As a result, stochasticity is introduced by sampling points from this normal distribution. Here, the encoder network ($q_\phi(z \mid x)$) compresses the input brain MR image x to a latent vector z (via sampling), and then z is reconstructed back to an image $\bar{x}$ similar to the original image, by a decoder network ($p_\theta(x \mid z)$). The objective function to maximize is as follows:

$$\Lambda(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x \mid z)\right] - D_{KL} \left(q_\phi(z \mid x) || p_\theta(z)\right) \tag{1}$$

where, the first term denotes the reconstruction loss or the negative expected log-likelihood and the second term represents the regularizer, i.e. Kullback-Leibler divergence between the distribution of encoder $q_\phi(z \mid x)$ and p(z). Furthermore, by adding the ϵ parameter (an auxiliary noise), the Reparameterization trick allows backpropagation to flow through the deterministic nodes, which is accomplished by:

$$\begin{aligned} \mathbf{z} &= \mu + \sigma \odot \epsilon \\ \epsilon &\sim \mathcal{N}(0, I) \end{aligned} \tag{2}$$

where μ and σ indicate the mean and standard deviation respectively, and ⊙ refers to the element-wise product. Once the VAE training process is completed, the brain MR images are sent to the trained encoder, and the latent vectors are obtained and stored along with their class labels.

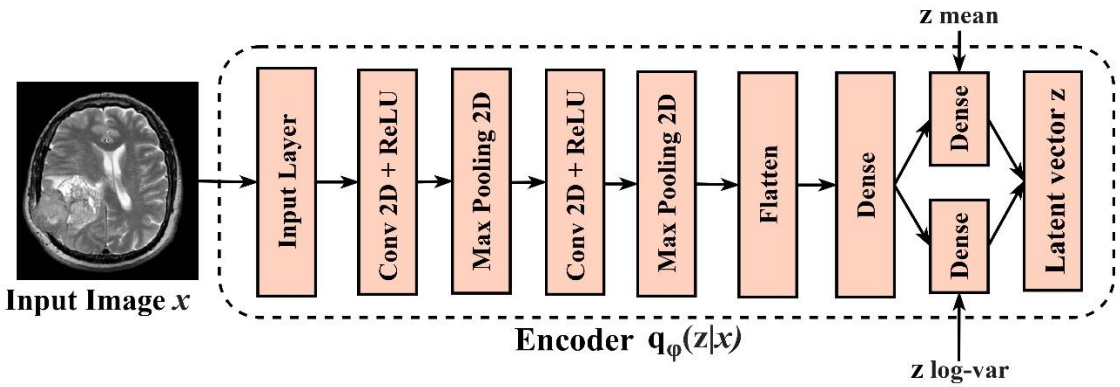

FIGURE 2: Encoder architecture

B. SEMI-SUPERVISED LEARNING FRAMEWORK

The second phase of the proposed architecture is the semi-supervised classification framework. As observed in Fig. 1, a Self-training approach has been implemented. On acquiring the feature embeddings, the data is segregated into two primary subsets, training and testing. For our experiments, 85% of the latent vectors with their class labels comprise the training set, and the testing set has 15% of the same. Furthermore, to utilize the self-trained SSL methodology, the training set is divided into labeled and unlabeled sets in various ratios to analyze and evaluate the proposed architecture.

Algorithm 1 Self Training Algorithm

Require: labeled Data $(X_{\mathcal{L}})$, Unlabeled Data $(X_{\mathcal{U}})$

repeat
  Train base classifier $(\mathcal{B})$ using $X_{\mathcal{L}}$
  $X_{\mathcal{P}} \leftarrow \mathcal{B}(X_{\mathcal{U}})$
  $X_{\mathcal{S}} \leftarrow$ Sort $X_{\mathcal{P}}$ based on confidence of $\mathcal{B}$
  $X_{\mathcal{K}} \leftarrow$ Select top-$\mathcal{K}$ confident samples from $X_{\mathcal{S}}$
  $X_{\mathcal{U}} \leftarrow X_{\mathcal{U}} \setminus X_{\mathcal{K}}$
  $X_{\mathcal{L}} \leftarrow X_{\mathcal{L}} \cup X_{\mathcal{K}}$
until Stopping criteria satisfied

Let the small labeled set be represented as $X_L = \{(x_i , y_i), i = 1, \ldots , n_l\}$, while the large unlabeled set be denoted as $X_U = \{x_i , i = 1, \ldots , n_u\}$, where x represents the latent vectors and y represents their class labels. Here, n_l and n_u signify the total number of labeled and unlabeled data, given $n_l << n_u$. The iterative process starts with training the initial base classifier (i.e. the learning algorithm) with the original labeled dataset X_L. Once trained, the classifier is used to predict the unlabeled set X_U. In this context, the predicted labels are referred to as 'pseudo-labels'. Next, as per the selection criterion, the most confident pseudo-labeled data samples (denoted by X_K) are added to the labeled set, i.e. $X_L \cup X_K$, and removed from the unlabeled set, i.e. $X_U \setminus X_K$. The classifier is then retrained with the updated labeled set. This process gets repeated numerous times until a predefined stopping criterion is met. It is important to note that at each iteration, the performance of the classifier model is evaluated using the testing set to monitor its behavior. The outline of the self-training algorithm that is used in the current study is reported in Algorithm (1). For our experiments, 4 different base classifiers have been considered., viz. Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Multi-layer Perceptron (MLP), and Support Vector Classifier (SVC). The models have been implemented using the scikit-learn [49] package.

SECTION V: EXPERIMENTAL ANALYSIS

This section describes our experimental framework. In the first subsection, the setup including the datasets used, the performance metrics used to access the framework, and the parameter setting is discussed. The second subsection is dedicated to the performance evaluation of the various base classifiers employed in the self-trained SSL framework. On obtaining the best-performing classifiers, the performances for different sizes of initial labeled data are compared in Section V-C. This helps us to understand the behavior of our model using different sizes of labeled data at the beginning of the iteration and indicates the optimal amount required to achieve the best results. The fourth subsection analyses the proposed framework with 2 fully supervised baseline models, viz. AdaBoost [50] and a standard deep neural network. In Section V-E, the proposed method is compared with state-of-the-art methods. Finally, the last section elaborates on the statistical significance test conducted to determine the efficacy of our framework.

A. SETUP

Datasets: To evaluate the effectiveness of our proposed framework, two brain MRI datasets have been considered for the task of brain tumor classification. The first dataset, viz. Brain Tumor Detection 2020 [51], is publicly available on Kaggle and has been referenced as BT1 throughout the paper for simplicity. It is a binary dataset comprising 3000 images, 1500 of which are images containing tumors, while the remaining 1500 images are without tumors. The second dataset [52], referred to as BT2, is a multi-class dataset, also taken from Kaggle. BT2 contains 7023 brain MR images categorized into four different classes, viz. glioma (1621 images), meningioma (1645 images), pituitary (1757 images), and normal (2000 images). Fig. 3 presents some sample images from BT1 and BT2 datasets.

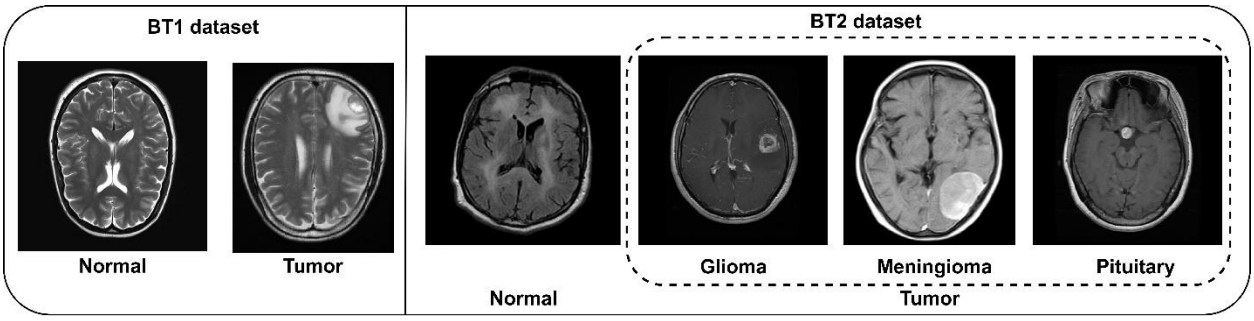

FIGURE 3: Sample brain MR images from the BT1 and BT2 dataset

Evaluation Measures: The proposed framework has been evaluated based on Accuracy, Precision, Recall, and F1-Score. In classification problems, these are some of the well-established [16] [17] metrics that are commonly used to evaluate a model. These performance metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive and false negative, respectively. In terms of our experiments, considering BT1 as an example, when the predicted and true classes are 'Tumor', then it is a true positive. On the other hand, if both the class labels are 'Normal', it is interpreted as a true negative. If the predicted class indicates 'Normal' and the true class label is 'Tumor', it is a false positive. Lastly, if the predicted class is 'Tumor' while the original class is 'Normal', it is regarded as a false negative.

Parameter Setting: All the brain MR images have been pre-processed and set to an image size of 128×128. Here, each image is a single-channel grayscale image. Before the training of the VAE, the images are normalized to [0, 1]. The dimensionality of the latent embeddings is chosen to be 96. The VAE has been trained using binary cross-entropy with Adam optimizer. The learning rate and the batch size are set to 5×10-4 and 256 respectively. Moreover, to combat overfitting, 'Early stopping' is implemented with a patience value of 3 while monitoring the total loss. As stated in Section III-B, the split between the labeled and unlabeled sets has been in various ratios. In other words, there have been 4 primary splits, viz. 5:95, 10:90, 15:85, and 20:80, that has been used to evaluate the self-trained SSL framework. In our case, to have a fair comparison, on reaching iteration 20, the learning process ceases. The selection criterion, as per Fig. 1, has been to consider the top 100 and 240 most confident pseudo-labeled data for BT1 and BT2 datasets, respectively. To avoid the class imbalance problem, in the case of the BT1 dataset, the 50 most confident pseudo-labeled data from each class are taken. Similarly, for BT2, the 60 most confident pseudo-labeled data from each of the 4 classes are picked.

Furthermore, to acquire the best performance from the classification algorithms, hyperparameter tuning has been conducted using GridSearch. For Logistic Regression, the hyper-parameters selected are the L2 penalty term, a tolerance of 1e-4, 'lbfgs' solver, and a maximum number of iterations of 1000. In the case of Gaussian Naïve Bayes, the variance smoothing value is 1e-09. At the same time, for the MLP classifier, the activation function set is ReLU, with Adam solver, the learning rate is chosen at a constant value of 0.001 and the maximum iteration is 1000. While for SVC, the regularization parameter is set to 1.0, the kernel chosen is 'rbf', with a tolerance of 0.001, and the probability estimates are enabled. In the case of our baseline models, the parameters for the AdaBoost classifier have been a learning rate of 1.0, the number of estimators has been chosen to be 50, and the algorithm is SAMME.R. Moreover, the base estimator is a Decision Tree classifier initialized with a maximum depth of 1. The second baseline model is a standard fully connected neural network with 3 hidden layers trained with Adam optimizer and binary cross-entropy as the loss function for 100 epochs. Meanwhile, for performance evaluation of the BT2 dataset, the macro-averaged Precision, Recall, and F1-Score have been considered. Our experiments were carried out using an Intel Core i5- 1035G1 CPU equipped with Intel UHD Graphics 620, 8 GB of RAM, Windows 10 Home 21H1, and TensorFlow 2.5.0.

B. COMPARISON AMONG THE DIFFERENT BASE CLASSIFIERS

This section focuses on the comparative study conducted on the BT1 and BT2 datasets, using 4 different base classifiers in the self-trained SSL framework. Fig. 4 and 5 illustrate the performance of the classifiers in terms of Accuracy, Precision, Recall, and F1-Score, for 20 iterations using BT1 and BT2 datasets, respectively. Each classifier is evaluated with the help of the testing set after every round of SSL training. Here, Iteration 0 is the state at which no pseudolabeled data is added to the labeled set. In addition, the initial size of the labeled set, in this particular set of experiments, is considered to be 10% of the training set. The rest 90% of the training set, with class labels discarded, forms the unlabeled dataset. Fig. 4a and 5a reveal that there has been a notable improvement in Accuracy scores in the case of the MLP classifier in both datasets. For GNB classifier in BT1, a steep decrement is observed in terms of Precision at iterations 0−2 (Fig. 4b). Around the same time, the Recall scores have increased considerably (Fig. 4c). Fig. 4d shows a striking improvement, in terms of F1-score, during iterations 17−20. For BT2, a gradual rise in the performance of the LR classifier has been recorded, in terms of Accuracy (Fig. 5a) and Precision (Fig. 5b). While, the SVC classifier has documented an overall

performance increment for Recall (Fig. 5c), and F1- Score (Fig. 5d), with few fluctuations. Between LR and MLP classifiers, 11.42% and 28.53% performance improvement is observed in terms of Accuracy over 20 iterations (Fig. 5a) implying that the MLP classifier is very well-suited to our proposed framework. Overall, the noteworthy improvement in the performance of the base classifiers implies that the low-dimensional embeddings, obtained from high-dimensional space, are meaningful and, with the help of the self-trained SSL set-up, good quality data from the unlabeled set is added to substantially improve the quality of the labeled set.

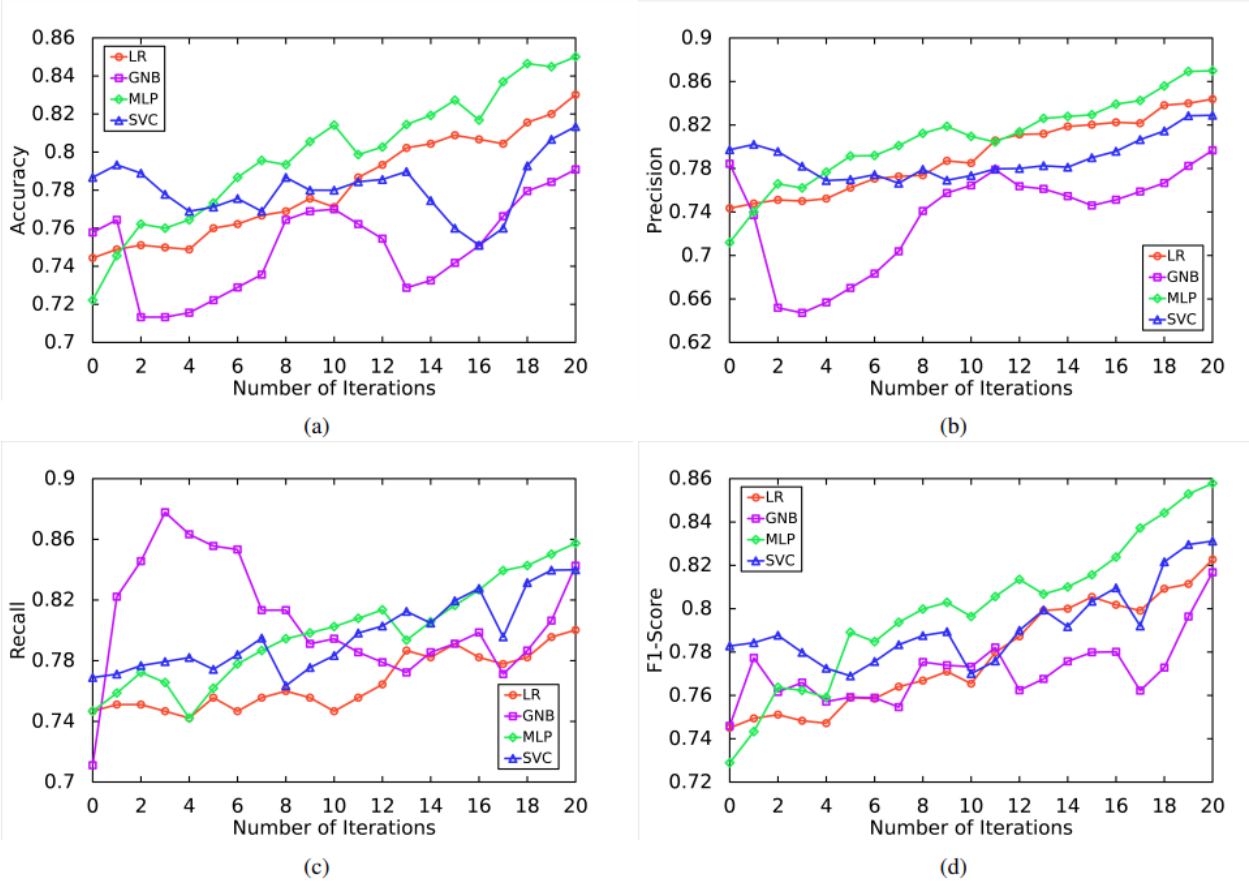

FIGURE 4: Performance evaluation among the different base classifiers using the BT1 dataset.

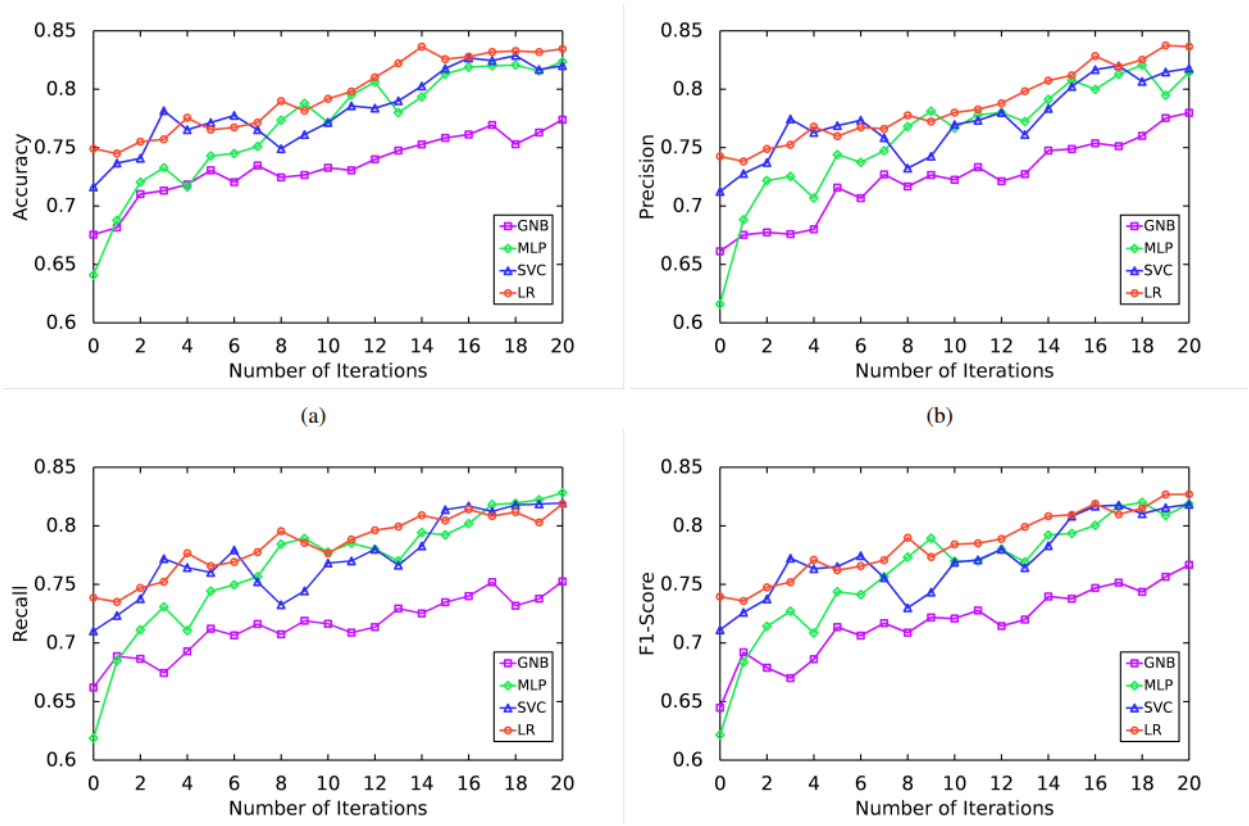

FIGURE 5: Performance evaluation among the different base classifiers using the BT2 dataset.

C. COMPARISON AMONG THE DIFFERENT SIZES OF THE INITIAL LABELED SET

To conduct a thorough analysis, Section V-C elaborates on the comparative study of the performances at various initial sizes of the labeled set, in terms of Accuracy, Precision, Recall, and F1-Score. As observed in Section V-B, the performance of the MLP classifier in our proposed framework has been quite remarkable. Hence, to conduct a fair assessment, in this set of experiments, it has been chosen as the base classifier. Similar to the previous experiments, the analysis presents the performance over 20 iterations with Iteration 0 as the state at which no pseudo-labeled data is added to the labeled set. Here, 4 different sizes of the initial labeled set have been considered, viz. 5%, 10%, 15%, and 20% of the training set. The rest of the data, with class labels removed, constitute the unlabeled set with 95%, 90%, 85%, and 80% of the training set, respectively. Fig. 6 and 7 exhibit the performances of the MLP classifier with 4

distinct initial labeled set sizes using BT1 and BT2 datasets, respectively. Here, Fig. 6 and 7 indicate that for a 20:80 split, the unlabeled set pool is exhausted at Iteration 19 and hence the process stops at this stage. It is quite evident from Fig. 6a and 7a that the original 10:90 split performs the best. On the contrary, the 5:95 (labeled: unlabeled) division records the most inferior performances over 20 iterations. In terms of Recall scores in the BT1 dataset, with 15% of the training data as the initial labeled set, significant fluctuations are noticed, which stabilize after Iteration 12 (Fig. 6c).

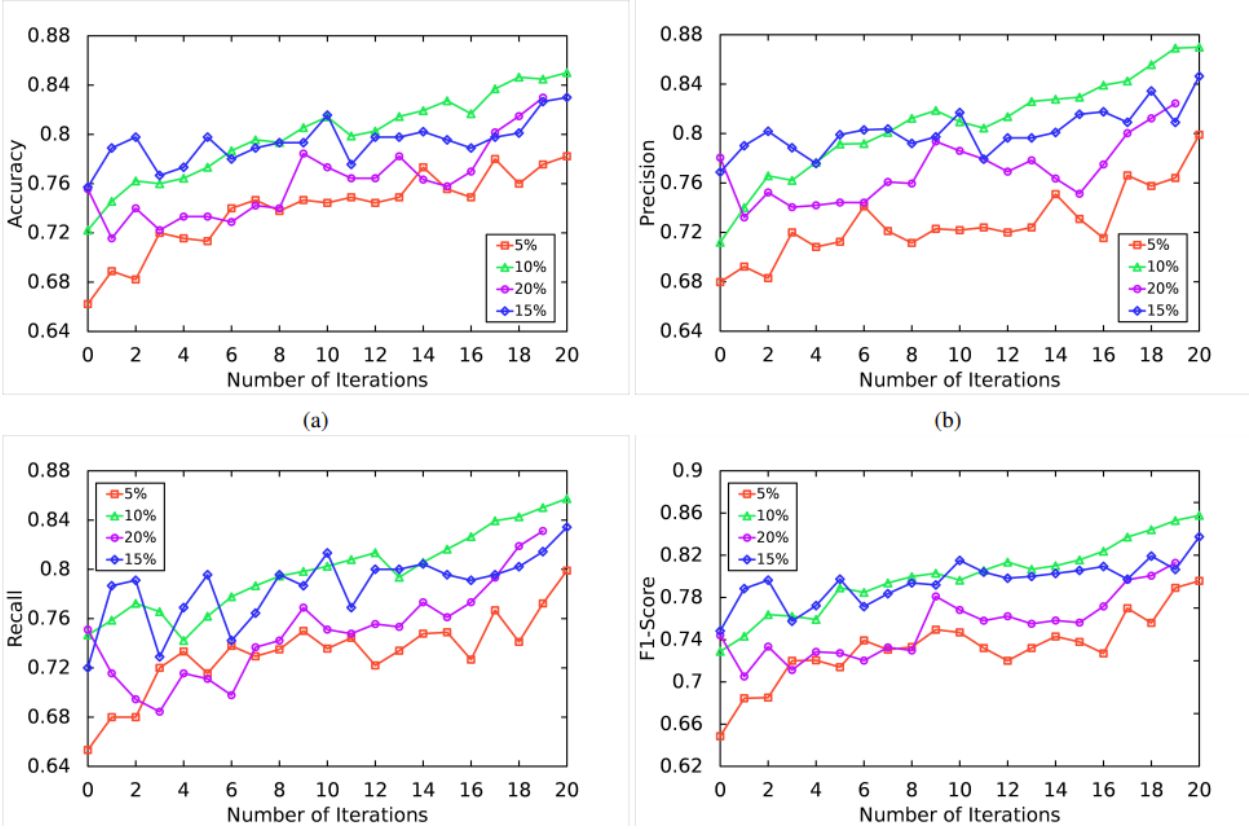

FIGURE 6: Performance evaluation among the different sizes of the initial labeled set using the BT1 dataset.

Surprisingly, the 20:80 split does not perform well over 20 iterations in comparison to the 10:90 and 15:85 splits, for both datasets. Moreover, the F1-score, being the harmonic mean of Precision and Recall, works as an important measure to evaluate the classifier performance at the different ratios. As observed in Fig. 6d, the performance of 10:90 and 15:85 for iterations 3-15 have been similar with few occasional differences. For the BT2 dataset, the performance in terms of Precision (Fig. 7b) and Recall (Fig. 7c), reveal that with much less amount of labeled data, the model fails to

improve dramatically when compared to other ratios. Fig. 7d records a sharp increase in MLP classifier in BT2 between Iterations 0-3. This trend is noticeable in the plots of all the other 3 performance metrics as well. Furthermore, in the case of the BT2 dataset, until Iteration 15 the performance of 15:85 is better than 10:90 in terms of Accuracy (Fig. 7a). Besides, for BT2, it is evident from the plots that after Iteration 6 there is not much improvement and the scores stagnate for a 20:80 split. Hence, overall, it can be inferred that the 10:90 split performs the best and exhibits the most improvement over 20 iterations in terms of Accuracy, Precision, Recall, and F1-score.

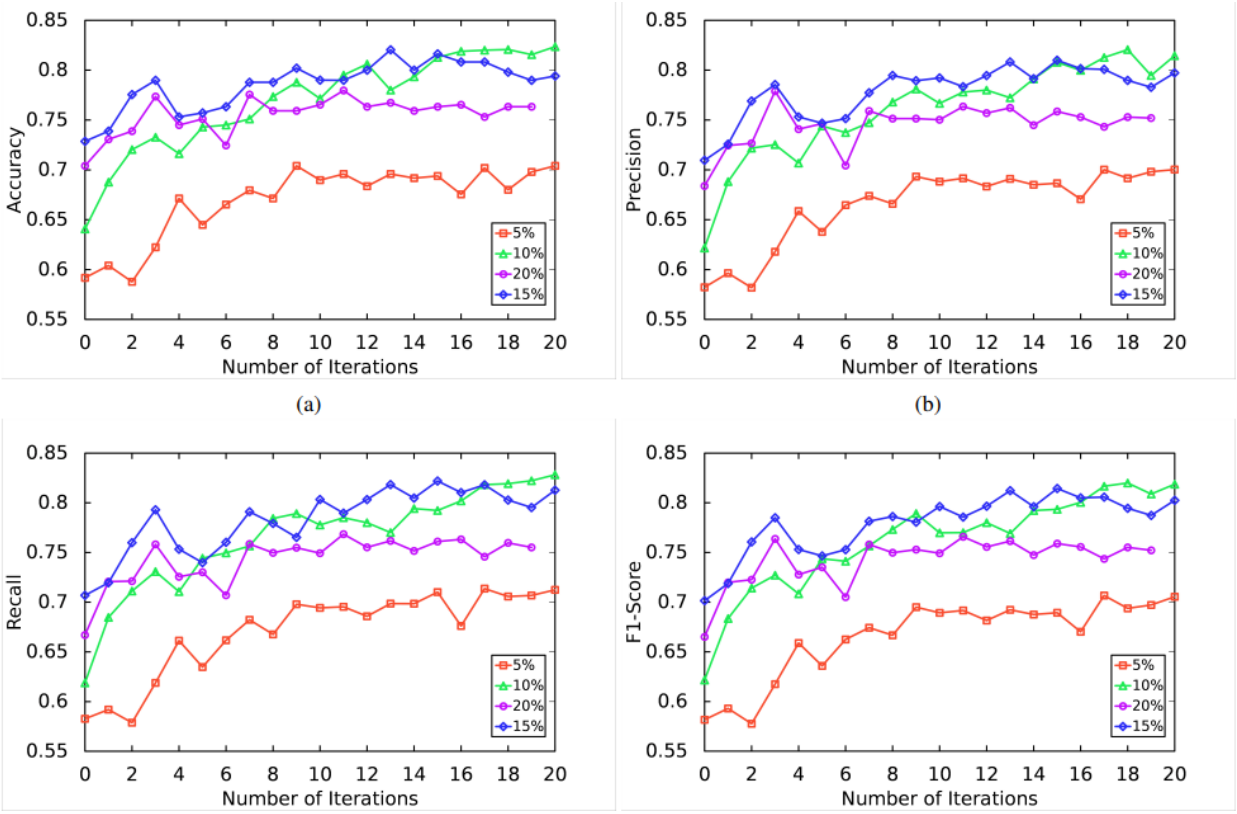

FIGURE 7: Performance evaluation among the different sizes of the initial labeled set using the BT2 dataset.

D. COMPARISON WITH BASELINE

The baselines considered for our experiments have been fully supervised AdaBoost [47] and feed-forward neural network (NN). Prior research [50], [51] on AdaBoost demonstrates that it is widely adopted in CAD literature. Being a meta-estimator, it has

been used to increase the efficiency of classifiers. At the same time, a standard neural network with 3 hidden layers has been used as the second baseline. Here, the training set (85% of the latent vector set with their class labels) has been used to train the baseline models. Meanwhile, the 15% testing set is used to evaluate them.

Performance metrics	Baseline	Proposed Approach with different base classifiers				
	AdaBoost	NN	LR	GNB	MLP	SVC
Accuracy	0.8484	0.8421	0.8302	0.7909	**0.8501**	0.8134
Precision	0.8593	0.8442	0.8438	0.7967	**0.8699**	0.8289
Recall	**0.8488**	0.8321	0.8002	**0.8426**	**0.8574**	0.8399
F1-Score	**0.8526**	0.8388	0.8227	0.8168	**0.8578**	0.8312

TABLE 1: Performance of the baseline in comparison to the proposed approach with different base classifiers, in the case of the BT1 dataset

Performance metrics	Baseline	Proposed Approach with different base classifiers				
	AdaBoost	NN	LR	GNB	MLP	SVC
Accuracy	0.8281	0.8122	**0.8345**	0.7739	0.8236	0.8201
Precision	**0.8399**	0.8241	**0.8365**	0.7798	0.8145	0.8179
Recall	**0.8206**	0.8154	0.8187	0.7525	**0.8282**	0.8194
F1-Score	**0.8281**	0.819	**0.827**	0.7667	0.8187	0.8183

TABLE 2: Performance of the baseline in comparison to the proposed approach with different base classifiers, in the case of the BT2 dataset

Tables 1 and 2 illustrate the performance comparison between the supervised baseline models and the proposed VAE-coupled self-trained SSL approach in terms of Accuracy, Precision, Recall, and F1-Score for BT1 and BT2 datasets. Further, in the case of the performance scores recorded for the respective classifiers of our proposed approach, they have been that of the 10:90 split and Iteration 20, which is the last iteration conducted to add pseudo-labeled data to the labeled set. It is quite evident from both the tables that our proposed self-trained SSL framework performs almost as well as the fully supervised baselines. With a limited amount of data in the case of SSL methodology (labeled set grows in size where the estimator teaches itself based on its

own predictions) can reach the performance of the supervised classification models, which are trained with 100% labeled set. The tables suggest that the MLP classifier for the BT1 dataset (Table 1), while LR and for the BT2 dataset (Table 2) in our algorithm have performed very well.

E. COMPARISON WITH STATE-OF-THE-ART

The proposed VAE-based semi-supervised latent space learning framework for brain tumor classification has been thoroughly analyzed in the previous sections. In the current section, a comparative analysis with state-of-the-art methods, already reported in the literature (Section II), has been presented. All the state-of-the-art approaches chosen for comparison have employed semi-supervised learning for brain tumor detection.

Method	Accuracy	Precision	Recall	F1-Score
Ge et al. [8]	0.8653	0.9102	0.7375	0.8147
Kamal et al. [40]	0.877	0.879	0.879	0.878
Albu et al. [39]	0.74	-	-	0.76
Azmi et al. [42]	0.7878	0.8003	-	-
Wan et al. [43]	0.9252	-	0.7309	0.7558
Tupe et al. [41]	0.8235	0.82	-	0.8
Proposed method	**0.8501**	**0.8699**	**0.8574**	**0.8578**

TABLE 3: Performance comparison with state-of-the-art methods.

Table 3 reports the performance comparison of the proposed framework with existing state-of-the-art methods. For the performance of the best model setup of our proposed strategy, MLP as the base classifier with the BT1 dataset has been reported. Here, the performance metric values of the proposed method are shown in bold. It can be observed from Table 3 that our framework has outperformed performances recorded in [39], [41], [42] in terms of Accuracy. Also, the performance of our model in terms of F1-Score, which is the harmonic mean of precision and recall, has been quite promising for the SSL task. In terms of recall, our method has obtained competitive performances for the study conducted in [40]. Hence, it can be inferred that the proposed framework is successful in predicting the labels of unlabeled latent vector data, with the performance reaching the current state-of-the-art approaches in terms of Accuracy, Precision, Recall, and F1-Score.

F. STATISTICAL ANALYSIS

To establish the effectiveness of our proposed framework, a statistical significance test has been conducted. Wilcoxon signed-rank test [52] at a significance level of 0.05 has been used to compare the classification models used in our approach. Here, the Null hypothesis holds that there is no significant mean difference between the performance indicators of the two base classifiers, whereas the alternative hypothesis argues that there is a significant difference. Each experiment has been run 30 times and average performance indicator values for individual classification algorithms have been computed. Tables 4 and 5 depict the P-values obtained from the non-parametric statistical rank test for all possible pairs of classifiers in terms of Accuracy. The cases where the null hypothesis has been rejected are in bold and it reveals that at a 95% level of confidence majority of the tests are statistically significant. Hence, this study proves that the performance obtained by various classifiers is not random and they are statistically sound.

	LR	GNB	MLP
GNB	**4.28E-03**	-	-
MLP	**1.15E-02**	**5.27E-03**	-
SVC	**2.99E-03**	2.12E-01	**3.18E-02**

TABLE 4: Wilcoxon signed-rank test statistic in terms of Accuracy using the BT1 dataset

	LR	GNB	MLP
GNB	**4.43E-02**	-	-
MLP	**8.30E-03**	1.58E-01	-
SVC	**9.17E-03**	**3.52E-02**	**1.17E-02**

TABLE 5: Wilcoxon signed-rank test statistic in terms of Accuracy using the BT2 dataset.

SECTION VI: CONCLUSION & FUTURE SCOPE

The current study proposed a semi-supervised latent space learning framework to efficiently detect brain tumors. The challenge of a lack of good-quality labeled MRI data is addressed by allowing the proposed model to be trained with a small amount of labeled data followed by a self-training-based training of the model using unlabeled data. To learn the best features from images, a VAE model is used to convert input MRI images to a latent vector form. All labeled and unlabeled images are then converted to latent vector form using the encoder of trained VAE. Various base classifiers are used to justify the ingenuity of the proposed framework. Experiments have revealed that the multilayer perceptron model has performed best as a base classifier in the case of brain tumor prediction as a binary classification task. Whereas, in the case of multiclass disease prediction, the logistic regression base classifier-based semi-supervised learning model performed best. The 10:90 split has exhibited the most improvement over 20 iterations in terms of performance metrics. Additionally, our proposed framework has demonstrated competitive performances for fully-supervised baselines and existing state-of-the-art semi-supervised brain tumor detection strategies. At the same time, the Wilcoxon signed-rank test conducted at a significance level of 0.05 reveals that the performance obtained by various classifiers is not random and they are statistically significant. Hence, it can be inferred that the extraction of features in terms of latent vectors can notably boost the semi-supervised learner performance, as demonstrated in our study. Future analyses can be focused on understanding other learning paradigms that use unlabeled data to improve classifier performance in brain tumor classification.

REFERENCES

[1] 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. Khanna et al., "A review on a deep learning perspective in brain cancer classification," Cancers, vol. 11, no. 1, p. 111, 2019.

[2] H. H. Sultan, N. M. Salem, and W. Al-Atabany, "Multi-classification of brain tumor images using deep neural network," IEEE access, vol. 7, pp. 69 215–69 225, 2019.

[3] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, "A deep learning model based on concatenation approach for the diagnosis of brain tumor," IEEE Access, vol. 8, pp. 55 135–55 144, 2020.

[4] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," Ieee Access, vol. 6, pp. 9375–9389, 2017.

[5] R. P. Mandal, D. Dutta, S. Bhattacharjee, and S. Chakraborty, "Water content prediction in smart agriculture of rural india using cnn and transfer learning approach," Intelligent Decision Support Systems for Smart City Applications, p. 167, 2023.

[6] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," Machine Learning, vol. 109, no. 2, pp. 373–440, 2020.

[7] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," Medical image analysis, vol. 54, pp. 280–296, 2019.

[8] C. Ge, I. Y.-H. Gu, A. S. Jakola, and J. Yang, "Deep semi-supervised learning for brain tumor classification," BMC Medical Imaging, vol. 20, no. 1, pp. 1–11, 2020.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[10] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," IEEE Transactions on Knowledge and Data Engineering, 2021.

[11] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," arXiv preprint arXiv:1712.08708, 2017.

[12] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, 2017, pp. 1133–1141.

[13] S. Bhattacharjee, S. Maity, R. Sen, and S. Chatterjee, "Class biased sarcasm detection using variational lstm autoencoder," in Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing. Springer, 2022, pp. 289–297.

[14] S. Maity, R. P. Mandal, S. Bhattacharjee, and S. Chatterjee, "Variational autoencoder-based imbalanced alzheimer detection using brain mri images," in Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing. Springer, 2022, pp. 165–178.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," arXiv preprint arXiv:1812.05069, 2018.

[17] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in international conference on machine learning. PMLR, 2019, pp. 4114–4124.
[18] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Unsupervised pathology detection in medical images using conditional variational autoencoders," International journal of computer assisted radiology and surgery, vol. 14, no. 3, pp. 451–461, 2019.
[19] H. Akrami, A. A. Joshi, J. Li, S. Aydore, and R. M. Leahy, "Brain lesion detection using a robust variational autoencoder and transfer learning," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020, pp. 786–790.
[20] R. Wei and A. Mahmood, "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey," Ieee Access, vol. 9, pp. 4939–4956, 2020.
[21] C. Vogelsanger and C. Federau, "Latent space analysis of vae and intro-vae applied to 3-dimensional mr brain volumes of multiple sclerosis, leukoencephalopathy, and healthy patients," arXiv preprint arXiv:2101.06772, 2021.
[22] S. Pálsson, S. Cerri, and K. Van Leemput, "Prediction of mgmt methylation status of glioblastoma using radiomics and latent space shape features," in International MICCAI Brainlesion Workshop. Springer, 2022, pp. 222–231.
[23] W. M. Salama and A. Shokry, "A novel framework for brain tumor detection based on convolutional variational generative models," Multimedia Tools and Applications, vol. 81, no. 12, pp. 16 441–16 454, 2022.
[24] N. Wang, C. Chen, Y. Xie, and L. Ma, "Brain tumor anomaly detection via latent regularized adversarial network," arXiv preprint arXiv:2007.04734, 2020. [25] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: a brief introduction," Frontiers of Computer Science, vol. 13, no. 4, pp. 669–676, 2019.
[26] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," arXiv preprint arXiv:2103.00550, 2021.
[27] S. Shorewala, A. Ashfaque, R. Sidharth, and U. Verma, "Weed density and distribution estimation for precision agriculture using semi-supervised learning," IEEE access, vol. 9, pp. 27 971–27 986, 2021.
[28] Z. Feng, G. Huang, and D. Chi, "Classification of the complex agricultural planting structure with a semi-supervised extreme learning machine framework," Remote Sensing, vol. 12, no. 22, p. 3708, 2020.
[29] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp. 253–260.
[30] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller, "Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning," IEEE Access, vol. 6, pp. 22 196–22 209, 2018.
[31] D. Y. Choi and B. C. Song, "Semi-supervised learning for continuous emotion recognition based on metric learning," IEEE Access, vol. 8, pp. 113 443–113 455, 2020.

[32] J. Chen, Z. Yang, and D. Yang, “Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification,” arXiv preprint arXiv:2004.12239, 2020.

[33] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” in 33rd annual meeting of the association for computational linguistics, 1995, pp. 189–196.

[34] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in Proceedings of the eleventh annual conference on Computational learning theory, 1998, pp. 92–100.

[35] X. Wang, D. Kihara, J. Luo, and G.-J. Qi, “Enaet: Self-trained ensemble autoencoding transformations for semi-supervised learning,” arXiv preprint arXiv:1911.09265, vol. 2, 2019.

[36] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 687– 10 698.

[37] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, “Deep virtual adversarial self-training with consistency regularization for semisupervised medical image classification,” Medical image analysis, vol. 70, p. 102010, 2021.

[38] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, “Semisupervised learning with deep generative models,” Advances in neural information processing systems, vol. 27, 2014.

[39] A. Albu, A. Enescu, and L. Malagò, “Tumor detection in brain mris by computing dissimilarities in the latent space of a variational autoencoder,” in Proceedings of the Northern Lights Deep Learning Workshop, vol. 1, 2020, pp. 6–6.

[40] I. M. Kamal and H. Bae, “Semi-supervised binary classification with latent distance learning,” arXiv preprint arXiv:2211.15153, 2022.

[41] P. Tupe-Waghmare, P. Malpure, K. Kotecha, M. Beniwal, V. Santosh, J. Saini, and M. Ingalhalikar, “Comprehensive genomic subtyping of glioma using semi-supervised multi-task deep learning on multimodal mri,” IEEE Access, vol. 9, pp. 167 900–167 910, 2021.

[42] R. Azmi, B. Pishgoo, N. Norozi, and S. Yeganeh, “Ensemble semisupervised frame-work for brain magnetic resonance imaging tissue segmentation,” Journal of medical signals and sensors, vol. 3, no. 2, p. 94, 2013.

[43] Z. Wan, Y. Dong, Z. Yu, H. Lv, and Z. Lv, “Semi-supervised support vector machine for digital twins based brain image fusion,” Frontiers in Neuroscience, p. 802, 2021.

[44] E. Puyol-Antón, C. Chen, J. R. Clough, B. Ruijsink, B. S. Sidhu, J. Gould, B. Porter, M. Elliott, V. Mehta, D. Rueckert et al., “Interpretable deep models for cardiac resynchronisation therapy response prediction,” in International Conference on Medical Image Computing and ComputerAssisted Intervention. Springer, 2020, pp. 284–293.

[45] S. Chatterjee, A. K. Das, J. Nayak, and D. Pelusi, “Improving facial emotion recognition using residual autoencoder coupled affinity based overlapping reduction,” Mathematics, vol. 10, no. 3, p. 406, 2022.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikitlearn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[47] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” Journal-Japanese Society For Artificial Intelligence, vol. 14, no. 771-780, p. 1612, 1999.

[48] A. Hamada, “Br35h :: Brain tumor detection 2020 dataset.” available online at: https://www.kaggle.com/datasets/ahmedhamada0/ brain-tumor-detection, last accessed on 01.07.2022.
[49] M. Nickparvar, “Brain tumor mri dataset.” available online at: https: //www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, last accessed on 01.07.2022.
[50] B. Mainak, K. Venkatanareshbabu, S. Luca, R. E. Damodar, C.-G. Elisa, M. R. Tato, N. Andrew et al., “State-of-the-art review on deep learning in medical imaging,” Frontiers in Bioscience-Landmark, vol. 24, no. 3, pp. 380–406, 2019.
[51] J. Hatwell, M. M. Gaber, and R. M. Atif Azad, “Ada-whips: explaining adaboost classification with applications in the health sciences,” BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1–25, 2020.
[52] R. F. Woolson, “Wilcoxon signed-rank test,” Wiley encyclopedia of clinical trials, pp. 1–3, 2007.