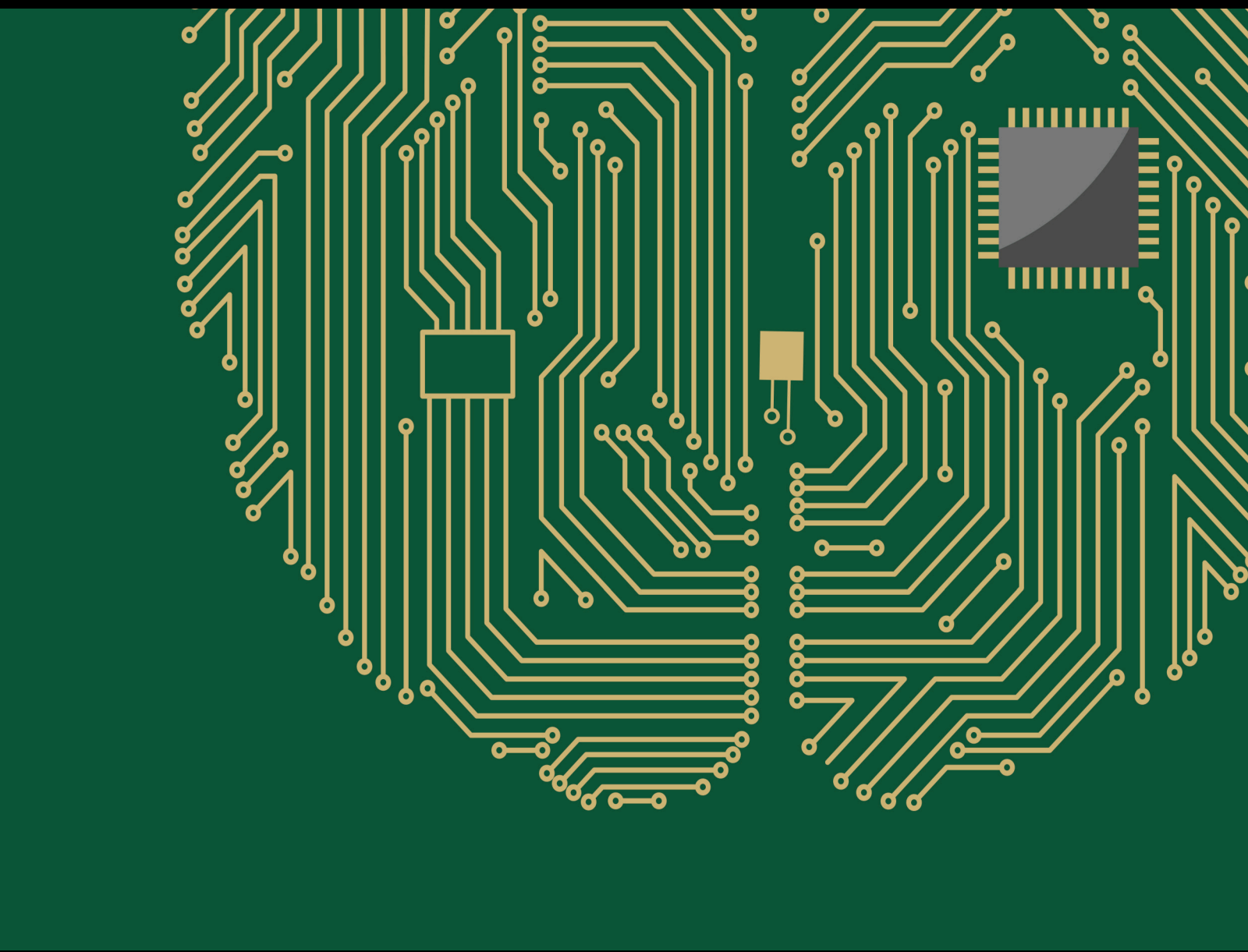
Computational Intelligence and Neuroscience

Biomedical Applications of Computer Vision using Artificial Intelligence

Lead Guest Editor: Vahid Rakhshan
Guest Editors: Alexandre Okano, Zhiyong Huang, Gianluca Castelnuovo, and Abrahão F. Baptista



Biomedical Applications of Computer Vision using Artificial Intelligence

Computational Intelligence and Neuroscience

Biomedical Applications of Computer Vision using Artificial Intelligence

Lead Guest Editor: Vahid Rakhshan
Guest Editors: Alexandre Okano, Zhiyong Huang, Gianluca Castelnuovo, and Abrahão F. Baptista

Contents

Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 9843574, 2 pages
https://doi.org/10.1155/2022/9843574

Editorial

Biomedical Applications of Computer Vision Using Artificial Intelligence

Vahid Rakhshan,[1] **Alexandre Hideki Okano**,[2] **Zhiyong Huang**,[3] **Gianluca Castelnuovo**,[4,5] **and Abrahão F. Baptista**[2]

[1]*Department of Cognitive Neuroscience, Institute for Cognitive Science Studies, Tehran, Iran*
[2]*Center for Mathematics, Computation and Cognition, Universidade Federal do ABC (UFABC), São Bernardo do Campo, Santo Andre 09606-045, Brazil*
[3]*Department of Computer Science, NUS School of Computing, National University of Singapore, Singapore*
[4]*Department of Psychology, Catholic University of Milan, Mila, Italy*
[5]*Istituto Auxologico Italiano IRCCS, Psychology Research Laboratory, Milan, Italy*

Correspondence should be addressed to Vahid Rakhshan; vahid.rakhshan@gmail.com

Received 5 January 2022; Accepted 5 January 2022; Published 26 January 2022

Computational neuroscience is concerned with simulating real neural systems to predict brain workings and disorders from subneuronal systems to network plasticity as hypotheses to be tested later in real neural tissues and hence to understand the principles governing them. Some ideas from this field can be used in artificial intelligence and other fields. Mimicking the central nervous system and by extension and creating various additional methods of computation such as artificial neural networks, machine learning, deep learning, or genetic algorithms have led to the artificial intelligence field which aims to solve given problems in a flexible, intelligent, and learnable way. The advent of these fields has numerous biomedical applications such as image processing and computer vision, machine learning, and deep learning for the assessment of imaging and signal datasets, disease diagnostic systems, expert systems to offer and optimize treatment planning, brain-computer interface, smart prosthetic limbs, and many others.

Image processing is a subfield of digital signal processing and a vast set of techniques used to enhance or manipulate digital images in order to make them more practical in different ways for different purposes. Computer vision is a field of computer science concerned with "understanding" images, videos, or 3D volumes by the computer through extracting desired features and attributes of images by various sets of algorithms and techniques.

This issue sought to publish select original and review articles on clinical and paraclinical applications of artificial intelligence and computational neuroscience in computer vision such as structural and functional brain imaging, histopathology, microbiology, surgery, and medical and dental radiography/tomography.

It published 6 articles: "Estimating Gender and Age from Brain Structural MRI of Children and Adolescents: A 3D Convolutional Neural Network Multitask Learning Model" by Mendes et al; "Denoising of 3D Brain MR Images with Parallel Residual Learning of Convolutional Neural Network Using Global and Local Feature Extraction" by Wu et al; "SGPNet: A Three-Dimensional Multitask Residual Framework for Segmentation and IDH Genotype Prediction of Gliomas" by Wang et al; "Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images Using Transfer Learning with EfficientNets" by Munien and Viriri; "A Semisupervised Learning Scheme with Self-Paced Learning for Classifying Breast Cancer Histopathological Images" by Asare et al; and "A Novel Bayesian Approach for EEG Source Localization" by Oikonomou and Kompatsiaris.

Data Availability

There are no data associated with this paper.

Conflicts of Interest

The editors declare that they do not have any conflicts of interest.

Vahid Rakhshan
Alexandre Hideki Okano
Zhiyong Huang
Gianluca Castelnuovo
Abrahão F. Baptista

Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 5550914, 12 pages
https://doi.org/10.1155/2021/5550914

Research Article

Estimating Gender and Age from Brain Structural MRI of Children and Adolescents: A 3D Convolutional Neural Network Multitask Learning Model

Sergio Leonardo Mendes,[1] Walter Hugo Lopez Pinaya,[2] Pedro Pan,[3] and João Ricardo Sato[1]

[1]*Center of Mathematics, Computing, and Cognition, Universidade Federal do ABC, Rua Arcturus no. 03, São Bernardo do Campo, SP 09606-070, Brazil*
[2]*Department of Biomedical Engineering, King's College London, London SE1 7EH, UK*
[3]*Escola Paulista de Medicina, Universidade Federal de São Paulo, R. Maj. Maragliano, 241 - Vila Mariana, São Paulo, SP 04017-030, Brazil*

Correspondence should be addressed to João Ricardo Sato; joao.sato@ufabc.edu.br

Received 28 January 2021; Revised 1 April 2021; Accepted 24 April 2021; Published 26 May 2021

Academic Editor: Vahid Rakhshan

Despite recent advances, assessing biological measurements for neuropsychiatric disorders is still a challenge, where confounding variables such as gender and age (as a proxy for neurodevelopment) play an important role. This study explores brain structural magnetic resonance imaging (sMRI) from two public data sets (ABIDE-II and ADHD-200) with healthy control (HC, $N = 894$), autism spectrum disorder (ASD, $N = 251$), and attention deficit hyperactivity disorder (ADHD, $N = 357$) individuals. We used gray and white matter preprocessed via voxel-based morphometry (VBM) to train a 3D convolutional neural network with a multitask learning strategy to estimate gender, age, and mental health status from structural brain differences. Gradient-based methods were employed to generate attention maps, providing clinically relevant identification of most representative brain regions for models' decision-making. This approach resulted in satisfactory predictions for gender and age. ADHD-200-trained models, evaluated in 10-fold cross-validation procedures on test set, obtained a mean absolute error (MAE) of 1.43 years (±0.22 SD) for age prediction and an area under the curve (AUC) of 0.85 (±0.04 SD) for gender classification. In out-of-sample validation, the best-performing ADHD-200 models satisfactorily predicted age (MAE = 1.57 years) and gender (AUC = 0.89) in the ABIDE-II data set. The models' accuracy was in line with the current state-of-the-art machine learning applications in neuroimaging. Key regions for models' accuracy were presented as a meaningful graphical output. New implementations, such as the use of VBM along with a 3D convolutional neural network multitask learning model and a brain imaging graphical output, reinforce the relevance of the proposed workflow.

1. Introduction

One of the current challenges faced by the mental health research field is to include biological measurements for the assessment of psychiatry disorders [1, 2]. Despite recent advances [3], psychopathology remains mainly assessed through clinical interviews [4, 5]. Investigations on neuroimaging biomarkers, particularly in youth, may help clinicians in the hard task of differentiating typical from atypical developmental trajectories.

Among several potential biomarkers, structural magnetic resonance imaging (sMRI) is a promising method to enhance identification and precise classification in psychiatry [6–8]. Moreover, characterizing atypical brain structures from sMRI is an important step for understanding the mechanisms and etiology of these disorders to tailor treatments [9]. Over the past few decades, dozens of studies have identified brain structural changes in ASD and ADHD [9–11]. However, the vast majority of these findings are inconclusive, possibly due to methodological issues such as

the use of small sample sizes, from a single study site, with little demographic variability (e.g., gender, age, or ethnicity) [9, 11]. These limitations have been recognized as a persistent source of bias in psychiatric classifications [12]. To achieve generalizable findings, one should employ large data samples, acquired from multiple sites/countries/scanners, including subjects with different ages, genders, ethnicities, and severity levels of psychiatry disorder [9, 11–13]. Fortunately, there are open data sets such as ABIDE-II and ADHD-200, which fit all these requirements.

Besides, most sMRI studies focused on traditional mass-univariate analytical methods, which are sensitive to gross and localized brain differences. These approaches, however, are not optimal for detecting subtle and spatially distributed neuroanatomical alterations, typically associated with psychiatric disorders [14, 15]. Therefore, machine learning techniques, such as deep learning networks, have shown interesting results in advancing group-level neuroimaging findings into individual-level clinically relevant classifications [16].

A specific deep learning network, called the convolutional neural network (CNN), revolutionized the computer vision area [17]. Regular CNNs use 2-dimensional images for their training process. This technical aspect, however, may cause loss of important data from the tridimensional (3D) structure of sMRI. A recent version of CNN, named CNN3D, overcomes this limitation by employing 3D images in its learning process, so it is an optimum candidate for sMRI applications. Recent studies, which used CNN to investigate psychopathologies, obtained better performance than the previously published literature [18–20]; however, none of these works employed a CNN3D trained with sMRI of youth to assess brain morphological features during neurodevelopment.

One downside of using deep learning models, such as CNN3D, is the low output interpretability, which sometimes provides little or no insight into the nature of the input data [14, 15]. To overcome this limitation, one can use a gradient-based algorithm such as SmoothGrad [21] to produce sensitivity voxel maps from input images that most contributed to models' decisions. Then, these attention maps can be intersected with a brain atlas such as AAL3 [22] to identify the top-focused brain regions of interest (ROIs) for the neural network decisions. This procedure may increase output interpretability and clinical relevance by showing brain ROIs with the greatest descriptive power for a given model prediction task. However, to date, few studies incorporated this approach. Moreover, integrating well-established sMRI processing techniques, such as voxel-based morphometry (VBM), into CNN3D training models seems to be appropriate to increase comparability to neuroimaging literature. VBM segments, aligns, and fits gray matter (GM) and white matter (WM) in a common spatial template, facilitating the hard task of comparing distinct clinical groups or gathering data for meta- or mega-analysis [23–26].

Different studies have contributed to the present knowledge on brain markers for psychiatric disorders, with several pieces of work assessing CNN3D [19, 20], multitask learning architecture [27, 28], and brain sMRI processed by VBM [9–11]. However, few studies have explored these methods jointly, particularly in large and heterogeneous data samples, to investigate biomarkers of neurodevelopment and psychiatric disorders across youth. The present study aims to evaluate a CNN3D model trained from ABIDE-II and ADHD-200 data sets to predict age (neurodevelopment), gender, and psychiatric disorder group (i.e., HC vs ASD or ADHD). We hypothesize that a CNN3D architecture, trained with 3D sMRI previously preprocessed by VBM, will detect complex patterns of morphological features in the human brain and allow correct classification of age, gender, and mental health status. Besides, we hypothesize that 3D saliency maps from trained models, generated via SmoothGrad [21], will provide identification of the brain's anatomical ROIs for each prediction task. These results could be intersected with 3D AAL3 brain atlas [22] and could be used to generate clinically relevant schematic representations of top-focused brain regions.

The current study evaluates the applicability of a workflow composed of carefully chosen methods and best practices to assess neurodevelopment from brain sMRI. First, the methods are described and justified in Section 2. Next, the achieved experimental results are presented in Section 3. Then, the results are discussed and compared to the related literature in Section 4. Finally, the conclusions are presented in Section 5.

2. Materials and Methods

2.1. Data Description. The data used in this study were obtained from two public data sets: Autism Brain Imaging Data Exchange II (ABIDE-II) and Attention Deficit Hyperactivity Disorder (ADHD-200). Both data sets can be downloaded from the NeuroImaging Tools & Resources Collaboratory Image Repository (NITRC-IR: https://www.nitrc.org/ir/). For this work, we used only one T1-weighted sMRI scan of each subject from the data sources. These images were collected from several locations in different countries: ABIDE-II includes 19 sites, and ADHD-200 includes 8 sites. Thus, the images' acquisition parameters vary due to different scanners' models and brands, ranging from 1.5T to 3T, each hosting a head coil from 8 to 32 channels. Detailed information and scanners' acquisition parameters can be retrieved from ABIDE-II (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html) and ADHD-200 (http://fcon_1000.projects.nitrc.org/indi/adhd200/) documentation. The data were collected and made public according to the responsibility and approval of the given local ethics of each project.

2.2. Subjects. Since we focus on neurodevelopmental processes in children and adolescents, we discarded subjects older than 20 years of age. Some individuals had more than one sMRI scan in the data set (collected from different scanning sessions). In these cases, only the first sMRI of each subject was considered. Data without information on gender, age, and psychiatric disorder (i.e., HC, ASD, or ADHD) were also discarded. Furthermore, each subject belonged exclusively to ABIDE-II or ADHD-200 data set (no subject

was in both). After applying these criteria, the sample for the present analysis and main demographic and phenotypic data are presented in Table 1 and Figure 1.

Individuals at different levels of the autism spectrum were grouped in the ASD label and, similarly, individuals with different subtypes of ADHD (inattention, hyperactivity, or combined) were grouped.

2.3. MRI Processing. The sMRI was processed using VBM [23] via the Statistical Parametric Mapping software (SPM12 v7771, from https://www.fil.ion.ucl.ac.uk/spm/software/). Briefly, VBM involves spatially normalizing all MRI images to the same stereotactic space, allowing extraction of different brain tissues from images partitioned with correction for nonuniform intensity variations [23]. In the past decades, VBM has been largely adopted in neuroimaging studies, such as the ones investigating ASD and ADHD [10]. The complete conceptual framework, methodology, and background behind the software are available in the Statistical Parametric Mapping book [29].

The data sets were processed using two batches of tasks (one batch for ABIDE-II and another for ADHD-200). Although the same procedures were applied to both data sets, we chose to process them in separate batches to ensure that each data set was completely independent. All the sMRI transformation steps were performed through the SPM12 software, following the VBM Tutorial [30].

First, sMRI data were spatially segmented to segregate GM and WM [24]. In this step, the skull, tissues, and artifacts outside the brain tissue are removed from the original image.

Second, the DARTEL algorithm [25] was applied to increase the accuracy of intersubject alignment. This transformation works by aligning GM among the images while simultaneously aligning WM during the generation of a template to which the data are iteratively aligned [26]. Third, the resulting files from the previous step were spatially normalized, Jacobian-scaled, and smoothed with a Gaussian full width at half maximum (FWHM) set to 8 mm to generate images in the Montreal Neurological Institute (MNI) coordinate system [31,32]. After these transformations, each sMRI scan produced two 3D matrices (one for GM and another for WM), with each voxel carrying the probable density of brain tissue at that location.

Finally, we loaded the previously transformed GM and WM via Python, through the SimpleITK library (https://simpleitk.org/) and applied a common mask assigning the value −1 to every background voxel (outside the brain). We chose to set the value −1 (instead of zero) to streamline the learning process of the models, due to the increase in the distance between background voxel values and brain voxel values with low tissue probability (close to zero). The brain matrices and their corresponding phenotypic data were saved in the TensorFlow record format (https://www.tensorflow.org/tutorials/load_data/tfrecord). This notation allows for better performance by storing data in binary linearly serialized files. As the data sets are still relatively large after the transformations (about 30 GB for both data sets), this step is important to read data efficiently during the model training phase.

2.4. Deep 3D Convolutional Neural Network Multitask Learning Architecture. The architecture of our model was designed to receive the previously transformed 3D brains as input for the neural network training. The input for training is a 5D matrix (composed of the number of examples in batch, voxel X-axis, voxel Y-axis, voxel Z-axis, brain tissues), where the brain tissue is a two-channel dimension composed of GM and WM. We considered only GM and WM to ensure that our models' predictions resulted from patterns directly related to differences in neurodevelopment. Therefore, the cerebrospinal fluid, the skull, and all the tissues outside the brain were discarded. That was also the reason why we did not use the complete unsegmented images. Moreover, we opted for feeding data through different channels to the model so that it had a facilitation signal to differentiate the patterns of GM (mostly neuronal nuclei) and WM (mostly axon bundles). As shown in Figure 2, the common model's body is composed of a sequence of interleaved layers of 3D convolution, batch normalization, and 3D max pooling, followed by dense and dropout layers. After the common model's body, we derived three output blocks, each composed of its own dense, batch normalization, and output layer. The output blocks are accountable to, respectively, predict gender, age, and psychiatric disorder (i.e., HC, ASD, or ADHD).

Inspired by the VGG16 network [33], we chose the ReLU activation function to provide nonlinearity [34] and used convolutional layers with receptive fields of $3 \times 3 \times 3$ pixels and max-pooling layers with $3 \times 3 \times 3$ pixel window and stride of $2 \times 2 \times 2$. To improve the network convergence, we added batch normalization [35] before convolutional and dense layers. To face overfitting problems, we included $L2$ kernel regularizers (with a coefficient equal to 1×10^{-3}) in all convolutional and dense layers and added a dropout [36] with a dropout rate of 0.5 right after the flattening of the last convolutional layer.

The loss chosen as the objective function to be minimized is expressed by the weighted sum of the loss of each output, where we opted for the Mean Squared Error for the age output and Binary Cross-Entropy for gender and diagnosis outputs. The loss weights (W_1, W_2, and W_3) were not tuned, remaining in the default values of the TensorFlow library (i.e., equal to 1). As the classification and regression tasks have different loss scales, the loss will be higher to the age estimation than to the classification tasks. That is, the training will tend to optimize more in the direction of the age estimation than in that of the classification tasks.

$$\begin{aligned} \text{objective}_{\text{loss}} = {} & W_1 * \text{mean\_squared\_error}\left(y_{\text{age}}, \widehat{y}_{\text{age}}\right), \\ & + W_2 * \text{binary\_crossentropy}\left(y_{\text{gender}}, \widehat{y}_{\text{gender}}\right), \\ & + W_3 * \text{binary\_crossentropy}\left(y_{\text{diagnosis}}, \widehat{y}_{\text{diagnosis}}\right). \end{aligned} \tag{1}$$

Our motivation for choosing a multitask learning architecture is the advantages produced by the learned features in the shared layers that are favored from the mechanisms of data amplification, attribute selection, eavesdropping, and representation bias [37]. In brief, this approach allows faster convergence and better generalization due to the extra information provided by the training signals of the related tasks [37].

Table 1: Subjects' demographic and phenotypic information.

Data set	N	Male (%)	Female (%)	Age, $y \pm$ SD	Age range, y	HC (%)	ASD (%)	ADHD (%)
ABIDE-II	580	73.8	26.2	12.12 ± 3.16	6.0–20.0	56.7	43.3%	—
ADHD-200	922	63.1	36.9	11.72 ± 2.99	7.1–19.9	61.3	—	38.7%

The number of subjects (N) is shown in numbers, while age is in years ± standard deviation and in range of minimum–maximum years of age.

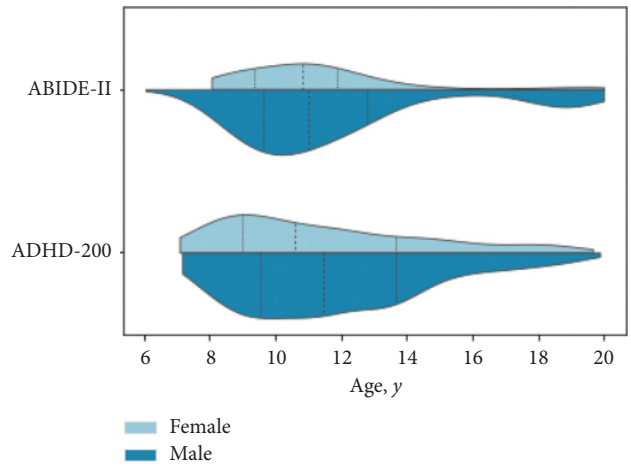


Figure 1: Subjects demographic distribution of ABIDE-II and ADHD-200 data sets. Vertical dotted lines show the quartiles. Ages are presented in years.

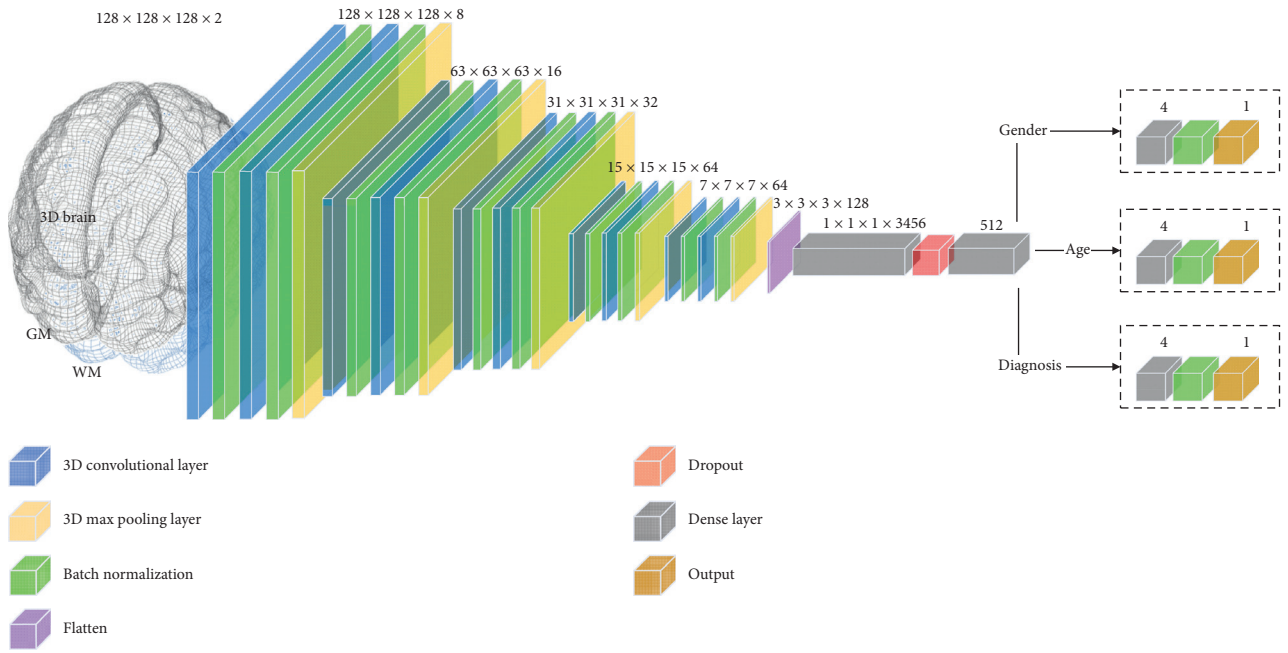


Figure 2: 3D convolutional neural network multitask learning model. The processing steps through the layers allow the extraction of increasingly complex brain features. While batch normalization allows faster network convergence, dropout plays an important role in increasing generalization. Due to the mechanisms of multitask learning architecture, such as data amplification and attribute selection, the shared features allow faster convergence and better generalization.

2.5. Model Tuning and Training. Despite our preferences for using an automated method for the tuning process (e.g., grid search or Bayesian optimization), which was already employed in other works [14, 15], the hundreds of hyperparameters combinations and the long time consumed by each training session made this strategy unfeasible. Instead, the tuning was carried out based on previous knowledge and mainly insights from the publications of the VGG16 network [33], batch normalization [35], and dropout [36].

To make better use of processing time and memory resources, we set the TensorFlow mixed-precision configuration to employ both 16-bit and 32-bit floating-point types

during the training phase (https://www.tensorflow.org/guide/mixed_precision). We also padded and trimmed the brain input matrix, which originally had the size of $121 \times 145 \times 121$ to $128 \times 128 \times 128$. This step only affected background voxels (outside the brain) whose values were all equal to −1. This procedure followed the TensorFlow performance guide, which states that feature matrices multiples of 8 or 128 should be used for best memory usage (https://cloud.google.com/tpu/docs/performance-guide).

To optimize the objective loss, we opted for a gradient-based method with adaptive learning rates named Adam [38]. The initial Adam's learning rate was set to 1×10^{-3}, and the exponential decay rates for the first and second estimate moments were, respectively, set to 0.9 and 0.999. The loss weights from the objective function were not tuned and may be further explored in an upcoming study.

For the training sessions, the batch size was set to 32 examples, which is the maximum size that fitted in memory. As our model deals with distinct target variables with different data distributions at the same time (i.e., age, gender, and mental health status), we opted to do not balance the classes at the batch level. Thus, the examples were just randomly shuffled before the batch splits. The number of epochs was set to 1000, and a custom early stopping technique was implemented to stop the training process every time there was no improvement of at least one of the output losses in the validation set for 75 consecutive epochs. Following this strategy, most (75%) training sessions ended after running from 150 to 300 epochs. Additionally, we employed a technique called model checkpoint. Therefore, at the end of each epoch, the model was evaluated against the validation set, and the best-performing model parameters for each task were saved. This strategy provides three model versions at the end of each training session: one performing better to predict gender, another performing better to predict age, and the last performing better to predict psychiatric disorder.

At a first glance, one may argue that it is counterintuitive to save different model versions from the same multitask learning based model. However, we found in our preliminary tests that this schema reduced the models' training until convergence by three times, when compared to the time spent to train three different single-task models. Additionally, this approach helped (1) to prevent overfitting, by saving the model weights at the optimum training point, and (2) to generate model versions trained to best extract the relevant features for its main task. We used the lowest loss of each output (i.e., *Mean Squared Error* for age prediction and *Binary Cross-Entropy* for gender and psychiatric disorder predictions) as the metrics to automatically save the best checkpoints.

2.6. Test Procedure. Each data set (ABIDE-II and ADHD-200) was stratified (i.e., balanced) by mental health status (i.e., HC, ASD, and ADHD), randomly shuffled, and split in a 10-fold cross-validation custom scheme. Accordingly, data is initially split into 10 partitions and, in every training round, 1 partition is chosen for the test set. Then, from the 9 remaining partitions, the first 8 are assigned to the training set and the last 1 is assigned to the validation set (see Figure S1 in the Supplementary Materials). This cross-validation scheme resulted in 10 training rounds for each data set. For each round, the corresponding training set was used to train the network. The remaining validation set was employed to automatically save the best-performing models through the previously described model checkpoint technique. The test sets were kept untouched until the models were fully trained so that the performance of the final models could be assessed on an unbiased and unexplored data set. This custom validation scheme takes advantage of the robustness of a nested (double) cross-validation while preserving the lower time consumption of a nonnested cross-validation scheme.

For each training round of the 10-fold cross-validation, we obtained three final trained models: (1) optimized for gender, (2) optimized for age, and (3) optimized for psychiatric disorder classification. These models were evaluated as follows:

(a) All models trained with ABIDE-II data were evaluated on their corresponding test set

(b) All models trained with ADHD-200 data were evaluated on their corresponding test set

(c) The best-performing model trained with ABIDE-II data to predict age was evaluated across the full ADHD-200 data set

(d) The best-performing model trained with ABIDE-II data to predict gender was evaluated across the full ADHD-200 data set

(e) The best-performing model trained with ADHD-200 to predict age was evaluated across the full ABIDE-II data set

(f) The best-performing model trained with ADHD-200 to predict gender was evaluated across the full ABIDE-II data set

The chosen metrics to evaluate the models' performance in the regression task of predicting age were *MAE (mean absolute error)*, *Pearson's correlation*, *P value* of the Pearson's correlation, and *R2-score* (also known as prediction R^2, cross-validation R^2 or q^2, which best assesses numerical accuracy for regression tasks [39]). For the tasks of predicting gender and psychiatric disorder, we used *precision* (specificity measure), *recall* (sensibility measure), *F1-score* (harmonic mean between precision and recall), and *AUC-ROC (area under the receiver operating characteristic curve)*. The *F1-score* was chosen (instead of the simple accuracy) due to its capability to evaluate unbalanced data better.

The use of unbalanced data for the gender and mental health status classifications can bias the models towards classifying minority cases as majorities [40]. To address this issue, we employed a ROC operating point selection that maximizes the harmonic mean between sensitivity and specificity [40]. That is, for each trained model, we use the validation data to find the cutoff value that best maximizes the balance between sensitivity and specificity.

The chosen cutoff value is then used to collect the metrics from the test data.

2.7. Model Interpretability. In general, artificial neural networks have been known for their low interpretability level, sometimes being labeled as a "black box" providing little or no insight into the nature of data [14, 15]. The explanation of image-based artificial neural networks remains a challenge in the healthcare domain. To address this issue, we employed an algorithm called SmoothGrad [21]. It produces a sensitivity map of the voxels that most contribute to the neural network decisions by measuring the impact that small perturbations applied to input images produce in the output gradients. Although SmoothGrad uses the same basic methodology as other algorithms, it has the advantage of producing sharpen results due to the strategy of applying different perturbations to the same input image. Moreover, it averages the resulting maps, producing a better smoothing effect [21]. The present study employed the SmoothGrad algorithm through the open-source library implementation called tf-keras-vis (available at https://pypi.org/project/tf-keras-vis).

As quoted in the original paper [21], the sensitivity map algorithms often produce signed values. Therefore, there is considerable ambiguity in how to convert these signed values to visualization colors, as the direction of the gradient is context-dependent. To solve this ambiguity, we opted for using the absolute values of the gradients, which has the potential of producing clearer pictures [41] and was also proposed by SmoothGrad authors [21]. During the attention maps generation, the noise level was set to 20%, and the number of samples (sample size) for each input image was set to 5. Although the SmoothGrad paper shows increasing definition in the produced maps as the sample size is incremented, the processing time for this task is directly proportional to the sample size. Therefore, higher sample size values proved to be unfeasible given our limited hardware resources. Furthermore, we verified in a preliminary test that setting sample size to 10 produced the same top ROIs as setting the chosen configuration of 5. As our models have three outputs, we had to set to zero all outputs that were not the ones chosen for measurement (e.g., while generating the age sensitivity map, we set the gender and psychiatric disorder outputs to zero).

Attention maps were generated for the final models of each of the 10 cross-validation folds from their corresponding test set. These maps were first averaged from their test set examples and then were normalized and averaged across all the 10 training rounds, resulting in an attention map for each task (i.e., predicting age, gender, or psychiatric disorder) and for each data set (ABIDE-II and ADHD-200). This strategy allowed for capturing common structural brain regions that are most descriptive for the models' decision-making in each task.

As the final generated maps have the same 3D shape of the input images (localized in the MNI space), we could identify the most predictive brain ROIs taking the intersection between the attention maps and the AAL3 3D brain atlas [22]. Finally, the maps were rendered in the MRIcron viewer (https://www.nitrc.org/projects/mricron) to provide more interpretable brain visualizations.

2.8. Experiments Setup. The sMRI processing steps were done through the software SPM12 v7771, *Python* v3.6.9, and TensorFlow v2.1.0, running on a local Linux desktop (CPU 3.2 GHz Octa Core, 32 GB ram). After the sMRI processing, the TFRecord files were uploaded to a Google Cloud storage bucket.

Our machine learning experiments were conducted using a Google Colab instance (https://colab.research.google.com/): CPU 2.3 GHz Dual Core, 12 GB ram, attached to a Cloud TPU v2 (180 teraflops/s speed and 64 GB ram), connected to the aforementioned storage bucket, Trough *Python* v3.6.9, and TensorFlow v2.3.

3. Results

The training and testing phases occurred successfully with adequate processing time for all models. Output metrics collected showed that CNN3D models were able to learn and predict age and gender with a high confidence level in both ABIDE-II ($MAE = 1.63 \pm 0.28$, $AUC = 0.82 \pm 0.06$) and ADHD-200 ($MAE = 1.43 \pm 0.22$, $AUC = 0.85 \pm 0.04$) data sets. For both age and gender predictions, models trained on ADHD-200 data had slightly higher performance than those trained on ABIDE-II, including when we evaluated the best-performing cross-validation models from one data set across the other distinct full data set ($MAE = 1.57$, $AUC = 0.89$ vs $MAE = 1.64$, $AUC = 0.79$).

For the age prediction, the ADHD-200 models evaluated in a 10-fold cross-validation scheme on the test set obtained an MAE (mean absolute error) of 1.43 years, reaching a mean Pearson correlation of 0.84 between the correct targets and the models' predictions and a mean R2-score (also known as prediction R^2, cross-validation R^2 or q^2) of 0.62. The best-performing model of the aforementioned cross-validation, which was trained with ADHD-200 data, achieved an MAE of 1.21 years on its corresponding test set, and when evaluated across the full ABIDE-II data set, it reached an MAE of 1.57 years and a Pearson correlation of 0.75 between targets and predictions (see Figure 3).

For gender prediction, the ADHD-200 models evaluated in a 10-fold cross-validation scheme on the test set obtained a mean AUC-ROC of 0.85, with precision = 0.84, recall = 0.81, and F1-score = 0.83. The best-performing model of the above-mentioned cross-validation, which was trained with ADHD-200 data, achieved an AUC-ROC of 0.91 on its corresponding test set, and when evaluated across the full ABIDE-II data set, it achieved an AUC-ROC of 0.89, with precision = 0.90, recall = 0.87, and F1-score = 0.89 (see Figure S2 in the Supplementary Materials).

For psychiatric disorder classifications, the models had poor learning, performing close to the random guessing. The ADHD-200 models evaluated in 10-fold cross-validation on the test set obtained a slightly better performance predicting ADHD (AUC-ROC = 0.61), while the models trained on

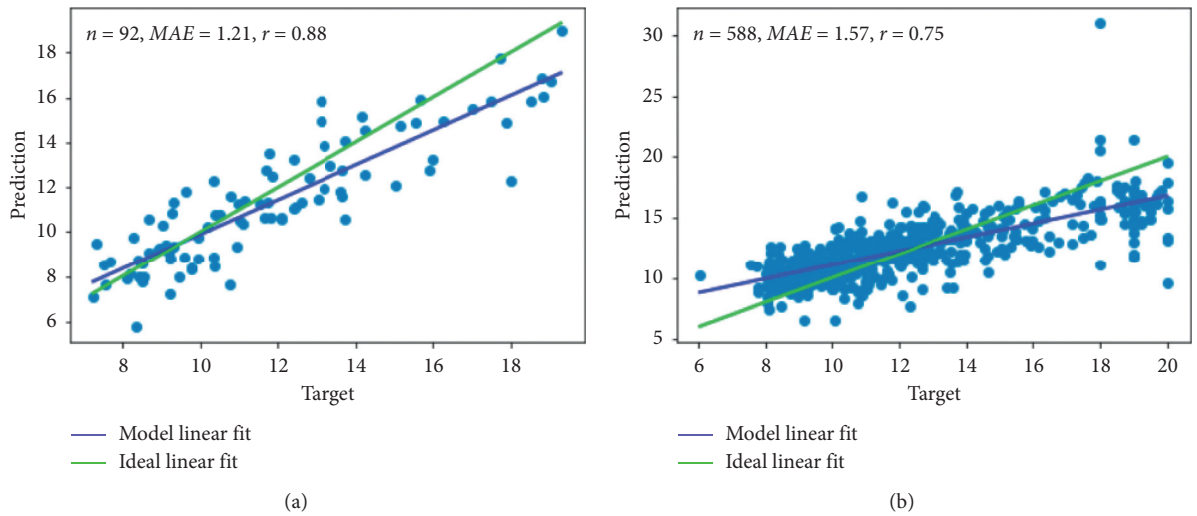


Figure 3: Scatter plots between predicted and target ages. (a) Ages prediction on the test set from the best-performing model of ADHD-200 10-fold cross-validation. (b) The same best-performing model, which was trained with ADHD-200 data, evaluated across the full ABIDE-II data set. Note. r: Pearson's correlation between predicted and target ages, MAE: the mean absolute error of the predictions, and n: the evaluated sample size.

ABIDE-II to predict ASD obtained a mean AUC-ROC = 0.54. All the evaluated metrics are presented in Table 2.

To access the statistical impact of the total brain volume on estimations, we calculated the AUC-ROC and Person's correlation (r), respectively, to gender and age concerning the sum of brain voxels from each subject. Thus, the ABIDE-II data ($N = 588$) yielded AUC-ROC = 0.76 and $r = 0.03$, while the ADHD-200 data ($N = 922$) resulted in AUC-ROC = 0.79 and $r < 0.001$. These results show that total brain volume is not related to age, while it may have influenced gender estimations. However, the focus of our work is the study of neurodevelopment, which is assessed mainly through age estimations.

The top 10 most representative ROIs from ADHD-200 models to classify gender are cingulate posterior gyrus (left and right), anteroventral thalamus (left and right), lateral posterior thalamus (right), mediodorsal lateral thalamus (right), mediodorsal medial thalamus (left and right), ventral anterior thalamus (right), and ventral lateral thalamus (right). In the ABIDE-II sample, the top 10 most representative ROIs comprised calcarine fissure (right), cingulate posterior gyrus (right), cerebellum lobe III (left), lingual gyrus (right), rolandic operculum (left), substantia nigra pars reticulata (left), pulvinar lateral thalamus (right), pulvinar medial thalamus (right), and vermis (lobes III and IV-V). The cingulate posterior gyrus (right) emerged as a top ROI on both ADHD-200 and ABIDE-II models for gender prediction.

Among age prediction models, the substantia nigra pars reticulata (left) arose in the top ROIs of both ADHD-200 and ABIDE-II models. ADHD-200 models retrieved the following regions as the top 10 ROIs: cingulate posterior gyrus (right), precentral gyrus (right), rolandic operculum (right), globus pallidus (left), substantia nigra pars reticulata (left), intralaminar thalamus (left), lateral geniculate thalamus (left), medial geniculate thalamus (left), pulvinar lateral thalamus (left), and vermis (lobes IV-V). ABIDE-II models top 10 focused ROIs comprised the following regions: the amygdala (right), middle cingulate (right), olfactory cortex (right), paracentral lobule (right), ventral tegmental area (right), vermis (lobes III and X), substantia nigra pars compacta (right), and substantia nigra pars reticulata (left and right). Interestingly, the vermis lobe III arose as a focused top 10 prediction ROI for gender and age in ABIDE-II models, and the vermis lobes IV-V emerged for both gender and age predictions in both samples. A compilation of the top-focused ROIs is depicted in Figure S3 in the Supplementary Materials.

As previously explained, model interpretability of artificial neural networks is sometimes challenging, which limits its applicability in clinical scenarios. Therefore, these models are deemed to be "black box," with little practical impact. However, we implemented a visualization approach to add to the models' interpretability. In Figure 4, we present an implementation of this procedure by adding the averaged gradients' attention maps as an overlayed layer of an MRIcron's brain template. It shows a practical example of visual outputs from artificial neural networks, where the top 10 predictive ROIs from gradients attention maps were accurately plotted in a clinically relevant representation of the brain.

4. Discussion

In this study, we transformed brain sMRI of youth via VBM, from large and heterogeneous data sets, and used the resultant GM and WM as input for training 3D's convolutional neural network with multitask learning models to predict age, gender, and psychiatric disorder. Then, the resultant trained models were used to map the top representative ROIs for the tasks of predicting age and gender. To achieve consistency and avoid biased results, we used a set of methods in line with the literature's best practices.

Table 2: Performance metrics of the test procedure.

Regression models	*n*	*MAE, y*	*r*	*P value*	*R2-scr*
Age: ABIDE-II 10-fold CV on test set	58	1.63 ± 0.28	0.76 ± 0.07	<0.001	0.54 ± 0.1
Age: ABIDE-II model on ADHD-200 full data	922	1.64	0.72	<0.001	0.50
Age: ADHD-200 10-fold CV on test set	92	**1.43** ± 0.22	0.84 ± 0.04	<0.001	0.62 ± 0.08
Age: ADHD-200 model on ABIDE-II full data	580	**1.57**	0.75	<0.001	0.56
Classification models	*n*	*Precision*	*Recall*	*F1-scr*	*AUC-ROC*
Gender: ABIDE-II, 10-fold CV on test set	58	0.87 ± 0.06	0.80 ± 0.08	0.83 ± 0.04	0.82 ± 0.06
Gender: ABIDE-II model on ADHD-200 full data	922	0.76	0.80	0.78	0.79
Gender: ADHD-200, 10-fold CV on test set	92	0.84 ± 0.03	0.81 ± 0.06	0.83 ± 0.03	**0.85** ± 0.04
Gender: ADHD-200 model on ABIDE-II full data	580	0.90	0.87	0.89	**0.89**
ASD: ABIDE-II, 10-fold CV on test set	58	0.46 ± 0.04	0.70 ± 0.18	0.55 ± 0.06	0.54 ± 0.06
ADHD: ADHD-200, 10-fold CV on test set	92	0.48 ± 0.07	0.55 ± 0.20	0.50 ± 0.11	**0.61** ± 0.07

The performance indicators from 10-fold cross-validation are presented in their averaged values ± standard deviation. The chosen model for the cross-data set evaluation is the best-performing model of 10-fold cross-validation. For the column titles, *r* is the Pearson's correlation between predicted and target ages, *n* is the sample size, and R2-scr is the prediction R^2 (also known as cross-validation R^2 or q^2). Values in bold are metrics of the best-performing trained models. ASD: autism spectrum disorder; ADHD: attention deficit hyperactivity disorder.

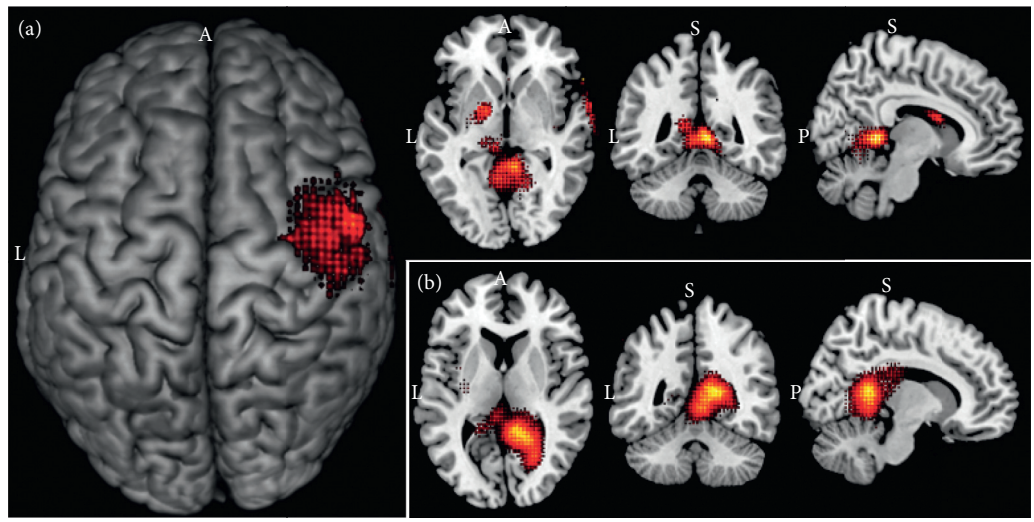


Figure 4: Top ROIs from gradients' attention maps perspective. (a) Top regions to predict the age by averaged attention of 10-fold ADHD-200 models. (b) Top regions to predict the gender by averaged attention of 10-fold ABIDE-II models. L: left, A: anterior, P: posterior, S: superior.

The ADHD-200-trained models had a slightly better performance than the models trained with ABIDE-II data, possibly because the first data set has higher homogeneity in data than the second [12]. The cross-data set evaluation proved the models' generalization capability to predict age and gender with high confidence even in unknown data sets with distinct confounding variables such as type of psychiatric disorder, scanner acquisition parameters, and subjects' distribution of age and gender.

To the best of our knowledge, the performance of our approach is in line with the state-of-the-art in brain aging detection, achieving an MAE = 1.43 years in 10-fold cross-validation on the test set. A study of Wang and coworkers [42] reached an MAE = 1.38 years in a subset from ADHD-200 with a similar age range to ours; however, their results were only based on healthy individuals, and their approach employed handcrafted feature extraction and selection based mainly on cortical thickness and curvatures. Another study, by Franke and colleagues [43], achieved an impressive MAE = 1.1 years in one of their test partitions and an MAE = 1.22 years from the averaged performance from all six test partitions. Unlike our work, Frank and coworkers employed a data set [44] acquired using a unified set of scanner parameters, from healthy subjects only, after rigorous filtering for dozens of confounding factors that could influence the healthy brain maturation during childhood and adolescence (i.e., individuals with preterm birth, alcohol or drug abuse during the gestational period, low IQ, and dozens of other confounding factors were excluded). Greater data uniformity, combined with smaller sample sizes, than that employed by us possibly provided good conditions so that both studies could achieve high accuracy [42,43], although it may have occurred at the cost of generalizability [12]. Different from our approach, these studies [42,43] employed a machine learning algorithm called relevance vector machine (RVM) [45], which is a Bayesian alternative to support vector machine. Therefore, RVM has the advantage of requiring less computational power than CNN3D.

Another study employed a CNN3D to predict age from brain sMRI in raw format versus sMRI processed by VBM. Cole and colleagues [46] achieved slightly better performance when they used VBM (MAE = 4.16 years) in comparison to raw sMRI (MAE = 4.66 years). However, they have only evaluated healthy subjects, with ages ranging from 18 to 90 years of age. Therefore, these differences do not allow a direct comparison of the model performance to our work. Additionally, unlike our study, Cole and coworkers [46] did not assess brain biomarkers (ROIs) from their model's predictions.

Although our approach presented a high capability to learn how to estimate age and gender, it did not perform well in classifying psychiatric disorders, achieving modest AUC-ROC and F1-score metrics when differentiating between HC, ASD, and ADHD. Therefore, the results show that our models were close to the random guessing for these tasks. Possibly, the underlying structural alterations from these conditions are subtle enough so that they are not efficiently detectable by CNN3D trained with sMRI from large and heterogeneous data sets. In psychiatric disorders, large and heterogeneous data samples tend to deliver high confidence and generalization power. However, at the same time, they tend to lead to low accuracies, which is an important limitation that possibly has also affected our main results [12]. Another source for investigation, in future work, is to evaluate the effect of tuning the weights from the objective loss function to prioritize the mental health status classification. The dynamic task prioritization for multitask learning [47] seems to be an interesting approach for this goal. This method proposes the dynamic adjustment of loss weights across the training process to prioritize the most difficult tasks.

The brain ROIs we identified (see Results) as being most representative for gender and age detection come in line with several distinct studies that reported these regions as being related to differentiation of gender, aging, or both [48–54].

For gender, Witte and coworkers [48] used Statistical Parametric Mapping to calculate GM volume differences between men and women, and among other statistically significant findings, they discovered that men had more GM than women in vermis, cerebellum, and right calcarine, while women had more GM than men in the lingual gyrus. Another study, by Menzler and colleagues [49], employed diffusion tensor MRI to discover microstructural differences between genders in the WM of the thalamus; Menzler and coworkers [49] also found differences in the cingulum confirming previous works, suggested that their findings were due to differences in myelination or glial cell morphometry, and stated that previous functional MRI studies found gender differences in thalamic activation during the processing of emotional stimuli or unpleasant linguistic information. Recent findings suggest that not only gender but also pubertal status may influence brain development [55]. Thus, the role of these features can be a source of further exploration in future work.

For age-related ROIs, Tomasi and Volkow [53] used functional MRI to evaluate the functional connectivity density (FCD) of networks concerning brain aging of healthy subjects and found that a long-range FCD in the default-mode network (DMN), which includes the posterior cingulate, decreased with age, while FCD in other two subcortical networks including thalamus and amygdala increased with age; more recently, an improved neuroanatomical model of DMN [56] not only included amygdala and thalamus in DMN but found that the thalamus has a centrality role in DMN. Another study used functional MRI [54] to find that in children the ventral tegmental area had lower connectivity to the amygdala and higher ventral tegmental area connectivity to the thalamus, globus pallidus, and vermis than in adults; this study [54] also revealed that in children the substantia nigra had higher connectivity to the amygdala, globus pallidus, and thalamus than in adults, and similarly the connectivity of language areas (including rolandic operculum) and middle cingulate was weaker with the ventral tegmental area than with substantia nigra for adults.

Taking it collectively, the morphological changes detected by our models and confirmed in other studies [48–54] are possibly related to the highly coordinated and sequenced events characterized by both progressive (myelination) and regressive (synaptic pruning) processes, which alter WM and GM volumes with different patterns for each gender, and are most dynamic from childhood to early adulthood [57].

These findings reinforce our hypothesis that CNN3D is able to detect complex brain morphological features, previously detectable by high-resolution diffusion tensor MRI and by functional MRI. Following Pinaya [15], while the standard mass-univariate techniques consider each brain structure as an independent unit, multivariate methods (such as the one we used) may be additionally based on interregional correlations leading individual regions to present high discriminative power due to two possible reasons: (a) a difference in volume/thickness between groups in that region; (b) a difference in the correlation between that region and other areas between groups. Therefore, discriminative brain networks are best interpreted as a spatially distributed pattern rather than as individual regions.

As our multitask learning architecture is optimized to perform all tasks at the same time (i.e., predicting gender, age, and psychiatric disorder), the learning process in the common model's body may favor the extraction of the brain features that are relevant to more than one task. On the other hand, each specific output block is exclusively optimized, selecting only the appropriate set of features that best help to accomplish its unique individual task.

Due to the complexity arising from the nonlinearity of artificial neural networks, our methods do not allow mapping the differences inside ROIs that are relevant to the models' decisions, that is, which patterns of increase/decrease in cortical volume of focused ROIs are accountable for a given model decision. Another limitation of the current study is that it does not explain the obtained performance results, that is, which methods are accountable for which performance improvements. Therefore, this topic is still open and can be further explored in future work.

Our approach was not sufficient to adequately classify ASD and ADHD. In contrast, the performance and generalization power achieved in predicting age (i.e., neurodevelopment) can pave the way for future work through the indirect estimation of psychiatric disorders. By training our model to predict the age of healthy individuals only (to be done), psychiatric conditions can be estimated by calculating the difference between the brain's predicted age and the individual's chronological age [46]. Increased brain predicted age has been detected in individuals progressing to Alzheimer's, in schizophrenia, in epilepsy, and Down's syndrome [58–61]. At the same time, decreased brain predicted age has been used to highlight the protective influences exerted by meditation, by increase in education levels, and by physical exercises [62, 63].

5. Conclusions

In conclusion, this study proved the ability of CNN3D models trained with GM and WM, processed via VBM, to accurately estimate age (i.e., neurodevelopment) and gender. Therefore, the achieved results endorse the hypothesis that our approach is able to detect complex brain patterns. Although the models were not able to efficiently differentiate between HC, ASD, and ADHD, the high performance and generalization power achieved in age estimation can pave the way for future work, through the indirect estimation of psychiatric disorders. The strategy of generating 3D brain saliency maps via SmoothGrad [21] and intersecting the results with the 3D AAL3 brain atlas [22] was successfully achieved. Therefore, it provided clinically relevant identification of most representative biomarkers (ROIs) during models' decisions and proved to be a viable alternative to deal with the well-known low interpretability problem of deep learning models. Finally, the results achieved by the presented approach reinforce the hypothesis that it can be successfully adapted to tackle a varying set of problems involving brain morphological alterations.

Data Availability

The data used in this study were obtained from two public data sets: Autism Brain Imaging Data Exchange II (ABIDE-II) and Attention Deficit Hyperactivity Disorder (ADHD-200). Both data sets can be downloaded from the Neuro-Imaging Tools & Resources Collaboratory Image Repository (NITRC-IR: https://www.nitrc.org/ir/). The data were collected and made publicly available according to the responsibility and approval of the given local ethics by each project. Detailed information for these data sets and their acquisition parameters can be retrieved from ABIDE-II (http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html) and ADHD-200 (http://fcon_1000.projects.nitrc.org/indi/adhd200/).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

JRS was supported by São Paulo Research Foundation (FAPESP) under Grants nos. 2018/21934-5 and 2018/04654–9. WHLP was supported by the Wellcome Flagship Programme (WT213038/Z/18/Z). The authors acknowledge the Autism Brain Imaging Data Exchange II (ABIDE-II) consortium and each of the twenty-one participating sites for sharing their data with the scientific community. The funding sources for the ABIDE-II data set are listed at http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html. The authors thank the Attention Deficit Hyperactivity Disorder (ADHD-200) consortium and each of the eight participating sites for sharing their data with the scientific community. The funding sources for the ADHD-200 data set are listed at http://fcon_1000.projects.nitrc.org/indi/adhd200/. The authors also acknowledge the Google Colaboratory (https://colab.research.google.com/).

Supplementary Materials

Figure S1 schematically demonstrates the adopted custom validation scheme, which takes advantage of the robustness of a nested cross-validation while preserving lower time consumption. Figure S2 displays the confusion matrices from the best-performing ADHD-200 model classifying gender on its test set and across the full ABIDE-II data set. Figure S3 presents the most representative brain regions to estimate age and gender from ADHD-200- and ABIDE-II-trained models. (*Supplementary Materials*)

References

[1] B. N. Cuthbert and T. R. Insel, "Toward new approaches to psychotic disorders: The NIMH research domain criteria project," *Schizophrenia Bulletin*, vol. 36, no. 6, pp. 1061-1062, 2010.

[2] C. Scarpazza, M. Ha, L. Baecker et al., "Translating research findings into clinical practice: a systematic and critical review of neuroimaging-based clinical tools for brain disorders," *Transl Psychiatry*, vol. 10, 2020.

[3] A. T. Drysdale, L. Grosenick, J. Downar et al., "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature Medicine*, vol. 23, no. 1, pp. 28–38, 2017.

[4] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, American Psychiatric Association, Washington, DC, USA, 2013.

[5] J. R. Sato, C. E. Biazoli, G. A. Salum et al., "Association between abnormal brain functional connectivity in children and psychopathology: a study based on graph theory and machine learning," *The World Journal of Biological Psychiatry*, vol. 19, no. 2, pp. 119–129, 2017.

[6] L. Schmaal, D. J. Veltman, D. J. Veltman et al., "Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group," *Molecular Psychiatry*, vol. 21, no. 6, pp. 806–812, 2016.

[7] L. Schmaal and D. P. Hibar, "Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA major depressive

disorder working group," *Molecular Psychiatry*, vol. 22, no. 6, pp. 900–909, 2017.

[8] A. Zugman, A. Harrewijn, E. M. Cardinale et al., "Mega-analysis methods in ENIGMA: The experience of the generalized anxiety disorder working group," *Human Brain Mapping*, vol. 41, pp. 1–23, 2020.

[9] A. M. Pagnozzi, E. Conti, S. Calderoni, J. Fripp, and S. E. Rose, "A systematic review of structural MRI biomarkers in autism spectrum disorder: a machine learning perspective," *International Journal of Developmental Neuroscience*, vol. 71, no. 1, pp. 68–82, 2018.

[10] S. Lukito, L. Norman, C. Carlisi et al., "Comparative meta-analyses of brain structural and functional abnormalities during cognitive control in attention-deficit/hyperactivity disorder and autism spectrum disorder," *Psychological Medicine*, vol. 50, no. 6, pp. 894–919, 2020.

[11] F. Samea, S. Soluki, V. Nejati et al., "Brain alterations in children/adolescents with ADHD revisited: a neuroimaging meta-analysis of 96 structural and functional studies," *Neuroscience & Biobehavioral Reviews*, vol. 100, pp. 1–8, 2019.

[12] H. G. Schnack and R. S. Kahn, "Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters," *Front Psychiatry*, vol. 7, 2016.

[13] A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz, "Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 4, no. 2, pp. 108–120, 2019.

[14] W. H. L. Pinaya, A. Gadelha, O. M. Doyle et al., "Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia," *Scientific Report*, vol. 6, 2016.

[15] W. H. L. Pinaya, A. Mechelli, and J. R. Sato, "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study," *Human Brain Mapping*, vol. 40, no. 3, pp. 944–954, 2018.

[16] D. Lei, W. H. L. Pinaya, J. Young et al., "Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual," *Human Brain Mapping*, vol. 41, no. 5, pp. 1119–1135, 2020.

[17] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] B. Sen, N. C. Borle, R. Greiner, and M. R. G. Brown, "A general prediction model for the detection of ADHD and Autism using structural and functional MRI," *PLoS One*, vol. 13, 2018.

[19] K. Oh, W. Kim, G. Shen et al., "Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization," *Schizophrenia Research*, vol. 212, pp. 186–195, 2019.

[20] M. A. Aghdam, A. Sharifi, and M. M. Pedram, "Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks," *Journal of Digital Imaging*, vol. 32, no. 6, pp. 899–918, 2019.

[21] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," 2017.

[22] E. T. Rolls, C. C. Huang, C. P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *Neuroimage*, vol. 206, 2020.

[23] J. Ashburner and K. J. Friston, "Voxel-based morphometry-the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[24] J. Ashburner and K. J. Friston, "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839–851, 2005.

[25] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[26] J. Ashburner and K. J. Friston, "Computing average shaped tissue probability templates," *Neuroimage*, vol. 45, pp. 333–341, 2008.

[27] S. Emrani, A. McGuirk, and W. Xiao, "Prognosis and diagnosis of Parkinson's disease using multi-task learning," in *Proceedings of the ACM SIGKDD International Conference Knowledge Discover Data Mining*, pp. 1457–1466, Halifax, Canada, August 2017.

[28] P. Cao, X. Shan, D. Zhao, M. Huang, and O. Zaiane, "Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease," *Pattern Recognition*, vol. 72, pp. 219–235, 2017.

[29] W. Penny, K. Friston, J. Ashburner, S. Kiebel, and T. Nichols, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Elsevier, Amsterdam, Netherlands, 2007.

[30] Ashburner J.. VBM Tutorial 2015:1–18.

[31] A. Mechelli, C. Price, K. Friston, and J. Ashburner, "Voxel-based morphometry of the human brain: methods and applications," *Current Medical Imaging Reviews*, vol. 1, no. 2, pp. 105–113, 2005.

[32] J. Ashburner, "Computational anatomy with the SPM software," *Magnetic Resonance Imaging*, vol. 27, no. 8, pp. 1163–1174, 2009.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 2015.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[35] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML 2015*, pp. 448–456, Lille, France, July 2015.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1858, 2014.

[37] R. Caruana, L. Pratt, and S. Thrun, *Multitask Learning*, Kluwer Academic Publishers, New York, NY, USA, 1997.

[38] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR) 2015*, San Diego, CA, USA, May 2015.

[39] D. Scheinost, S. Noble, C. Horien et al., "Ten simple rules for predictive modeling of individual differences in neuroimaging," *Neuroimage*, vol. 193, pp. 35–45, 2019.

[40] B. Song, G. Zhang, W. Zhu, and Z. Liang, "ROC operating point selection for classification of imbalanced data with application to computer-aided polyp detection in CT colonography," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 1, pp. 79–89, 2014.

[41] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[42] J. Wang, W. Li, W. Miao, D. Dai, J. Hua, and H. He, "Age estimation using cortical surface pattern combining thickness with curvatures," *Medical & Biological Engineering & Computing*, vol. 52, no. 4, pp. 331–341, 2014.

[43] K. Franke, E. Luders, A. May, M. Wilke, and C. Gaser, "Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI," *Neuroimage*, vol. 63, no. 3, pp. 1305–1312, 2012.

[44] A. C. Evans, "The NIH MRI study of normal brain development," *Neuroimage*, vol. 30, no. 1, pp. 184–202, 2006.
[45] M. E. Tipping, "The relevance vector machine," *Neural Information Processing Systems*, vol. 22, pp. 653–658, 2000.
[46] J. H. Cole, R. P. K. Poudel, D. Tsagkrasoulis et al., "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," *Neuroimage*, vol. 163, pp. 115–124, 2017.
[47] M. Guo, A. Haque, D. A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic Task prioritization for multitask learning," *Lecture Notes in Computer Science (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*, vol. 220, 2018.
[48] A. V. Witte, M. Savli, A. Holik, S. Kasper, and R. Lanzenberger, "Regional sex differences in grey matter volume are associated with sex hormones in the young adult human brain," *Neuroimage*, vol. 49, no. 2, pp. 1205–1212, 2010.
[49] K. Menzler, M. Belke, E. Wehrmann et al., "Men and women are different: diffusion tensor imaging reveals sexual dimorphism in the microstructure of the thalamus, corpus callosum and cingulum," *Neuroimage*, vol. 54, no. 4, pp. 2557–2562, 2011.
[50] K. Keller and V. Menon, "Gender differences in the functional and structural neuroanatomy of mathematical cognition," *Neuroimage*, vol. 47, no. 1, pp. 342–352, 2009.
[51] P. Jacques, F. Dolcos, and R. Cabeza, "Effects of aging on functional connectivity of the amygdala during negative evaluation: a network analysis of fMRI data," *Neurobiology of Aging*, vol. 31, pp. 315–327, 2010.
[52] J. Velísková and S. L. Moshé, "Sexual dimorphism and developmental regulation of substantia nigra function," *Annals of Neurology*, vol. 50, pp. 596–601, 2001.
[53] D. Tomasi and N. D. Volkow, "Aging and functional brain networks," *Molecular Psychiatry*, vol. 17, no. 5, pp. 549–558, 2012.
[54] D. Tomasi and N. D. Volkow, "Functional connectivity of substantia nigra and ventral tegmental area: maturation during adolescence and effects of ADHD," *Cerebral Cortex*, vol. 24, no. 4, pp. 935–944, 2012.
[55] Z. Gracia-Tabuenca, M. B. Moreno, F. A. Barrios, and S. Alcauter, "Development of the brain functional connectome follows puberty-dependent nonlinear Trajectories," *Neuroimage*, vol. 13, Article ID 117769, 2021.
[56] P. N. Alves, C. Foulon, V. Karolis et al., "An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings," *Communications Biology*, vol. 2, 2019.
[57] T. J. Silk and A. G. Wood, "Lessons about neurodevelopment from anatomical magnetic resonance imaging," 2011.
[58] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, and H. Sauer, "Brain AGE in mild cognitive impaired patients: predicting the conversion to alzheimer's disease," *PLoS One*, vol. 8, 2013.
[59] N. Koutsouleris, C. Davatzikos, S. Borgwardt et al., "Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders," *Schizophrenia Bulletin*, vol. 40, no. 5, pp. 1140–1153, 2014.
[60] H. R. Pardoe, J. H. Cole, K. Blackmon, T. Thesen, and R. Kuzniecky, "Structural brain changes in medically refractory focal epilepsy resemble premature brain aging," *Epilepsy Research*, vol. 133, pp. 28–32, 2017.
[61] J. H. Cole, T. Annus, L. R. Wilson et al., "Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline," *Neurobiology of Aging*, vol. 56, pp. 41–49, 2017.
[62] J. Steffener, C. Habeck, D. O'Shea, Q. Razlighi, L. Bherer, and Y. Stern, "Differences between chronological and brain age are related to education and self-reported physical activity," *Neurobiology of Aging*, vol. 40, pp. 138–144, 2016.
[63] E. Luders, N. Cherbuin, and C. Gaser, "Estimating brain age using high-resolution pattern recognition: younger brains in long-term meditation practitioners," *Neuroimage*, vol. 134, pp. 508–513, 2016.

Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 5577956, 18 pages
https://doi.org/10.1155/2021/5577956

Research Article

Denoising of 3D Brain MR Images with Parallel Residual Learning of Convolutional Neural Network Using Global and Local Feature Extraction

Liang Wu,[1] **Shunbo Hu**,[2] **and Changchun Liu**[1]

[1]*School of Control Science and Engineering, Shandong University, Jinan 250061, China*
[2]*School of Information Science and Engineering, Linyi University, Linyi 276005, China*

Correspondence should be addressed to Shunbo Hu; 84961103@qq.com and Changchun Liu; changchunliu@sdu.edu.cn

Received 23 January 2021; Revised 15 April 2021; Accepted 21 April 2021; Published 4 May 2021

Academic Editor: Vahid Rakhshan

Magnetic resonance (MR) images often suffer from random noise pollution during image acquisition and transmission, which impairs disease diagnosis by doctors or automated systems. In recent years, many noise removal algorithms with impressive performances have been proposed. In this work, inspired by the idea of deep learning, we propose a denoising method named 3D-Parallel-RicianNet, which will combine global and local information to remove noise in MR images. Specifically, we introduce a powerful dilated convolution residual (DCR) module to expand the receptive field of the network and to avoid the loss of global features. Then, to extract more local information and reduce the computational complexity, we design the depthwise separable convolution residual (DSCR) module to learn the channel and position information in the image, which not only reduces parameters dramatically but also improves the local denoising performance. In addition, a parallel network is constructed by fusing the features extracted from each DCR module and DSCR module to improve the efficiency and reduce the complexity for training a denoising model. Finally, a reconstruction (REC) module aims to construct the clean image through the obtained noise deviation and the given noisy image. Due to the lack of ground-truth images in the real MR dataset, the performance of the proposed model was tested qualitatively and quantitatively on one simulated T1-weighted MR image dataset and then expanded to four real datasets. The experimental results show that the proposed 3D-Parallel-RicianNet network achieves performance superior to that of several state-of-the-art methods in terms of the peak signal-to-noise ratio, structural similarity index, and entropy metric. In particular, our method demonstrates powerful abilities in both noise suppression and structure preservation.

1. Introduction

Medical image information is playing an increasingly important role in disease diagnosis. However, during the image acquisition process, due to the improper actions of patients or staff, strong random noise will inevitably be generated. This noise not only reduces the resolution of the image but also affects the precision of clinician diagnosis [1, 2].

At present, popular magnetic resonance (MR) imaging technology is commonly used as a medical imaging technology for visualizing human tissues and organs. It does not pose any radiation hazard, unlike CT imaging [3], and it achieves multiaspect, multiparameter, and high-contrast-resolution images without bone artifacts. However, the random noise will affect the inspection quality in clinical diagnosis, as well as image processing and analysis tasks such as image segmentation, registration, and visualization. Hence, solving the problem of MR image denoising is critical.

The purpose of image denoising is to remove background noise and retain valuable information [4]. Many conventional filtering techniques are often used, such as Wiener filtering [5], bilateral filtering [6], and total variation filtering [7]. Yang and Fei proposed a multiscale wavelet denoising method based on the Radon transform to denoise MR images [8]. Phophalia et al. mitigated the problem of medical image denoising by using rough set theory (RST) [9]. Awate and Whitaker devised a Bayesian denoising

method and verified it on diffusion-weighted MR images [10]. Satheesh et al. developed an MR image denoising algorithm using the contourlet transform, which achieved a higher peak signal-to-noise ratio than the wavelet transform [11]. Zhang et al. used an improved singular value decomposition method to denoise simulated and real 3D images. The experimental results showed that their method was superior to the existing denoising methods [12]. Leal et al. presented a method based on sparse representations and singular value decomposition (SVD) for nonlocally denoising MR images. This method prevents blurring, artifacts, and residual noise [13]. In addition, by extending the local region to a nonlocal scheme, the nonlocal means (NLM) strategy was used for MR image denoising [14–16]. Gautam et al. proposed a novel denoising technique for MR images based on the advanced NLM method with nonsubsampled shearlet transform (NSST) [17]. Kanoun et al. proposed an enhanced NLM filter using the Kolmogorov-Smirnov (KS) distance. The experimental results provided excellent noise reduction and image-detail preservation [18].

In recent years, the explosive development of deep learning has suggested a new methodology for image denoising. It can use multiple convolution filters to automatically extract features, with large receptive fields, to reconstruct high-resolution images. In [19], the authors used the self-encoder to train the image features of different resolutions to achieve adaptive denoising. Zhang et al. exploited denoising convolutional neural networks (DnCNNs) for Gaussian noise removal and achieved excellent performance by using residual learning strategy [20]. Cherukuri et al. applied a deep learning network that leveraged the prior spatial structure of images to reconstruct high-resolution images [21]. Manjo′n et al. proposed a novel automatic MR image denoising method by combining a convolutional neural network (CNN) with a traditional filter [22].

Deep learning-based denoising methods can grasp richer contextual information in large regions to improve performance. With very deep architectures, it can expand the receptive field of the network to capture more global contextual information over large image regions. Liu et al. utilized the multiscale fusion convolution network (MFCN) to perform super-resolution reconstruction of MR images [23]. Pham et al. used a deep 3D CNN model with residual learning to reconstruct MR images [24]. Their model exploited a very deep architecture with a large receptive field to acquire a powerful learning ability. Jiang et al. described a multichannel denoising convolutional neural network (MCDnCNN) that directly learned the process of denoising and performed experiments on simulation and real MR data [25]. In [26], Ran et al. suggested a residual encoder-decoder Wasserstein generated countermeasure network (RED-WGAN) for MR image denoising. Hong et al. designed a spatial attention mechanism to obtain the area of interest in MR images, which made use of the multilevel structure and boosted the expressive ability of the network [27]. Tripathi and Bag proposed a novel CNN for MR image denoising. The proposed model consisted of multiple convolutions that captured different image features while separating inherent noise [28]. Li et al. designed a progressive network learning strategy by fitting the distribution of pixel-level and feature-level intensities. Their experimental results demonstrated the great potential of the proposed network [29]. Gregory et al. created HydraNet, a multibranch deep neural network architecture that learned to denoise MR images at a multitude of noise levels, and proved the superiority of the network on denoising complex noise distributions compared to some deep learning-based methods [30]. Aetesam and Maji proposed a neural framework for MR images denoising, using an ensemble-based residual learning strategy. High metric value and high-quality visual results were obtained in both synthetic and real noisy datasets [31].

In the above reported deep learning denoising tasks, the depth and width of the networks were often increased to capture more contextual information. However, these methods introduced a number of parameters, which made it difficult to train the denoising models. Some of the methods learned the Rician noise distribution solely by stacking convolution layers, which easily overlooked much local information and led to unsatisfactory denoising results at some key local anatomical positions.

To address the above shortcomings, this work proposes a novel network termed 3D-Parallel-RicianNet that is used to remove the noise of MR images. First, to expand the receptive field without introducing more parameters, we design a dilated convolution residual (DCR) module and use it to build a subnetwork (DCRNet) that can extract global information by cascading. Then a depthwise separable convolution residual (DSCR) module is designed and used to construct a subnetwork (DSCRNet) to extract local information. Finally, the features of each module of DCRNet and DSCRNet are merged and cascaded to obtain full-scale mappings between image appearances and noise deviation.

The main contributions of this work are summarized as follows:

(1) DCRNet expands the receptive field to extract rich context information through cascading DCR modules, which capture the real Rician distribution in the global area

(2) DSCRNet uses the DSCR module to focus on the local area of the image and effectively removes local anatomical noise. Each DSCR module of this subnetwork is added to the output part of each DCR module of the corresponding DCRNet

(3) The 3D-Parallel-RicianNet uses a residual learning mechanism to prevent vanishing and exploding gradient problems

The remainder of this work is organized as follows. In Section 2, we describe the proposed denoising networks and loss function. Then, in Section 3, we present the experimental tests of our approach on synthetic and real MR noisy data. Additionally, a comparison of our method with state-of-the-art algorithms is provided. Finally, in

Section 4, we discuss our conclusions and give future directions.

2. Materials and Methods

2.1. Noise Reduction Model. MR magnitude image is corrupted by independent Gaussian distribution noise in the real part and the imaginary part of images [32–34]. Previous studies suggest that the probability distribution of noisy MR image pixel intensity can be represented as a Rician distribution [35, 36]. Deep learning can ignore the physical process and model this procedure corruption by learning from the samples [26]. Hence, the MR image degradation model with noise can be described as

$$Y = X + \delta(Y), \tag{1}$$

where Y is the noisy MR image, X is the noise-free image, and $\delta(Y)$ is the deviation between X and Y influenced by the Rician distribution. According to equation (1), $\delta(Y)$ can be expressed as $(Y - X)$, so it was employed to train a residual mapping $f(Y;\Theta) \approx \delta(Y)$, and we can obtain $X \approx Y - f(Y;\Theta)$. Figure 1 shows that the probability density distribution (PDF) of noisy MR images varies in global and local regions. It can be seen from the top left image that the noise reduces the quality of the MR image and blurs the boundaries of some tissue structures, which results in increased difficulty in recognizing the image details. Liu et al. pointed out that the PDFs of Rician noise vary spatially in different anatomical regions of brain MR images [37]. Hence, the nonlinear mappings between image appearances and Rician distributions vary in global and local regions. Based on this conclusion, we propose the 3D-Parallel-RicianNet MR image denoising model, which combines the global and local feature information on global regions and local regions.

2.2. DCR Module for Global Feature Representation. It is known that context information is important to reconstruct corrupted pixels for image denoising. Specifically, it is a common way to capture more global context information by expanding the receptive field [38]. In the reported deep learning denoising tasks, increasing the depth and width of the deep networks can enlarge the receptive field. However, the width-adding methods may produce more parameters, which results in overfitting of the network. The depth-adding methods may lead to vanishing gradients when the depth of the network is enormous.

To solve these problems, dilated convolutions have been developed [39]. The dilation rate of the convolution kernel can be controlled to obtain receptive fields of different sizes, as shown in Figure 2. The size of the receptive field, v, is denoted as

$$v = \left(\left(k_{\text{size}} - 1\right) \times (R - 1) + k_{\text{size}}\right)^d, \tag{2}$$

where k_{size} is the size of the filter, R is the dilated rate, and d is the dimension (2 or 3) of the image. The receptive field of the convolution operation can be expanded by setting different R. This creates a tradeoff between increasing the depth and width of CNNs. In [40], Peng proposed dilated residual networks with symmetric skip connection (DSNet). The experiments demonstrated that the model was more feasible for the task of image denoising, especially for Gaussian noise. Zhang et al. proposed a dual-domain multiscale CNN (DMCNN) for JPEG artifacts based on dilated convolution. This also proved that dilated convolution had advantages in restoring image quality [41].

In this study, we construct the DCR module as one component of our 3D-Parallel-RicianNet. It exploits dilated convolutions to extract global features, as shown in Figure 3. The DCR module consists of dilated convolution, residual learning, batch normalization (BN), and leaky rectified linear unit (LeakReLU). Residual learning fundamentally breaks the symmetry of the network, thereby improving the ability of the representation network. By setting the BN layer, the generalization ability of the network is improved. Due to the problem of vanishing gradients using the ReLU activation function, we use LeakReLU as the activation function of the network. The input and output of a two-level dilated convolution are briefly connected to construct a DCR module.

2.3. DSCR Module for Local Feature Representation. It is very important to recover the local fine details in image denoising. When some local features are not well extracted, the local denoising effect will be degraded. Recently, depthwise separable convolution (DSConv) has been used in many advanced neural networks, such as Xception [42], MobileNets [43], and MobileNets2 [44], to replace the standard convolutional layer, aiming to reduce CNN computational cost and to extract local features [45].

DSConv consists of two parts: depthwise convolution and pointwise convolution. As shown in Figure 4, the depthwise convolution acts on each input channel separately, to exact local features, followed by a pointwise convolution that uses $1 \times 1/1 \times 1 \times 1$ convolution to weight the features among channels at every point. Hence, this would efficiently extract the local features among different channels. The input feature map is $I = \{I_1, I_2, \ldots, I_n\}$. First, using depthwise convolutions with n filters $K = \{K_1, K_2, \ldots, K_n\}$, an intermediate result $J = \{J_1, J_2, \ldots, J_n\}$ is produced, which is then processed into the output feature map $O = \{O_1, O_2, \ldots, O_m\}$ by means of the pointwise convolutions using m filters $k = \{k_1, k_2, \ldots, k_m\}$.

DSConv can extract local delicate features of the image by considering the information of the position and channel separately. Imamura et al. designed a denoising network for hyperspectral images using DSConv and demonstrated its ability to realize efficient restoration [46]. The advantage of DSConv is that it reduces the number of network parameters and the computational complexity in convolution operations [42–44].

The model designed by using dilated convolution can restore the image quality globally [40, 41] but can easily ignore local information. To solve this problem, inspired by DSConv, we extend the technique to the DSCR module to extract the local information of the MR images, as shown in

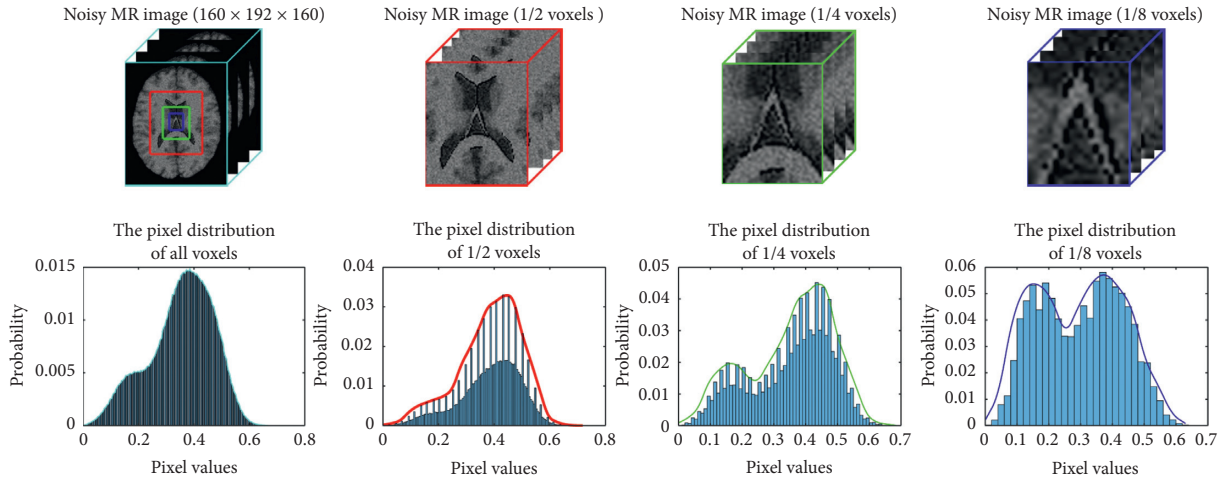


Figure 1: PDFs in different sizes of subwindows. The height of each vertical bar is the proportion of the corresponding pixel value. The lines represent the fitted pixel distribution. Top row from left to right: a 3D T1-weighted MR image with 7% Rician noise, a 1/2-voxel T1-weighted MR image, a 1/4-voxel T1-weighted MR image, and a 1/8-voxel T1-weighted MR image. Bottom row PDFs are of intensity in the corresponding voxels. As shown, the PDF (1/2 voxels) within the red region tends to be similar for the whole image. However, the PDFs in the small local green region (1/4 voxels) and local blue region (1/8 voxels) are different from those in the global region.

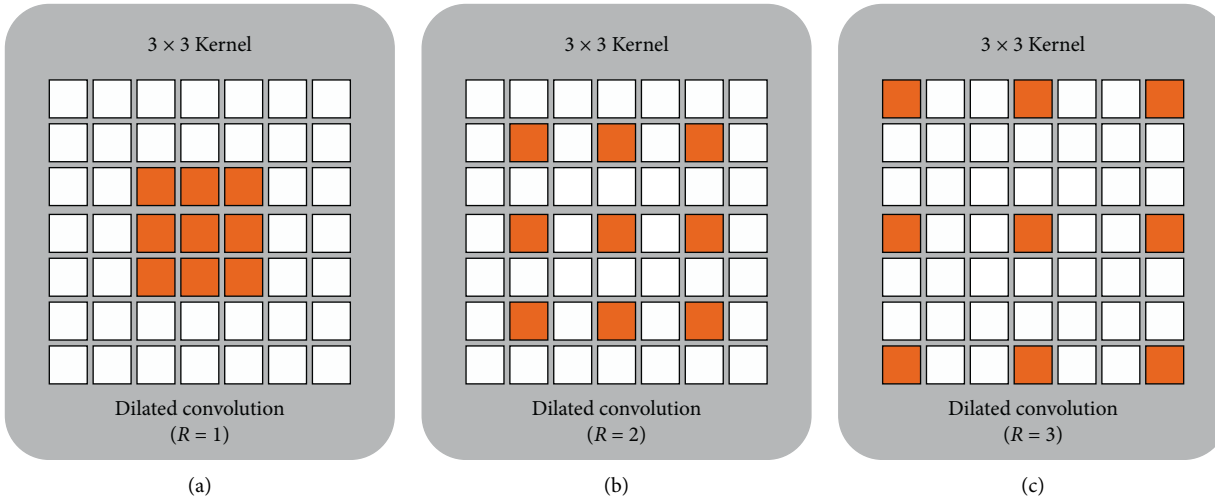


Figure 2: Dilated convolution. (a) The receptive field of the convolution kernel of size 3 × 3 pixels, which covers a 3 × 3 subregion in the image through a convolution operation. (b) The receptive field of the convolution kernel with an R of 2. (c) The receptive field of the convolution kernel with an R of 3.

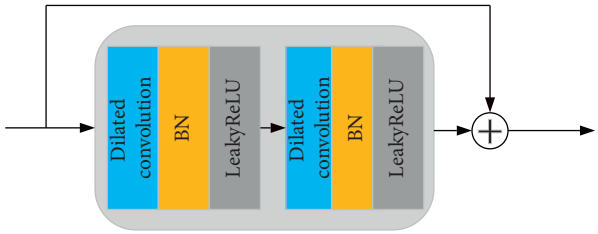


Figure 3: The structure of the DCR module.

Figure 5. We utilize the residual strategic idea and take the depthwise separable convolutions as the main construction module. On the one hand, we design two continuous depthwise separable convolutions with the BN layer after each convolution layer to improve the generalization ability of the network. On the other hand, we use another depthwise separable convolution to shortcut the module to prevent vanishing gradients.

2.4. The Proposed 3D-Parallel-RicianNet Model. The proposed 3D-Parallel-RicianNet framework consists of a global feature extraction network DCRNet, a local feature extraction network DSCRNet, and a reconstruction (REC) module. Under this framework, the pipeline of MR image denoising is composed of three major steps (see Figure 6). First, we apply DCRNet and DSCRNet to extract the global features and local features, respectively. Then, we fuse the

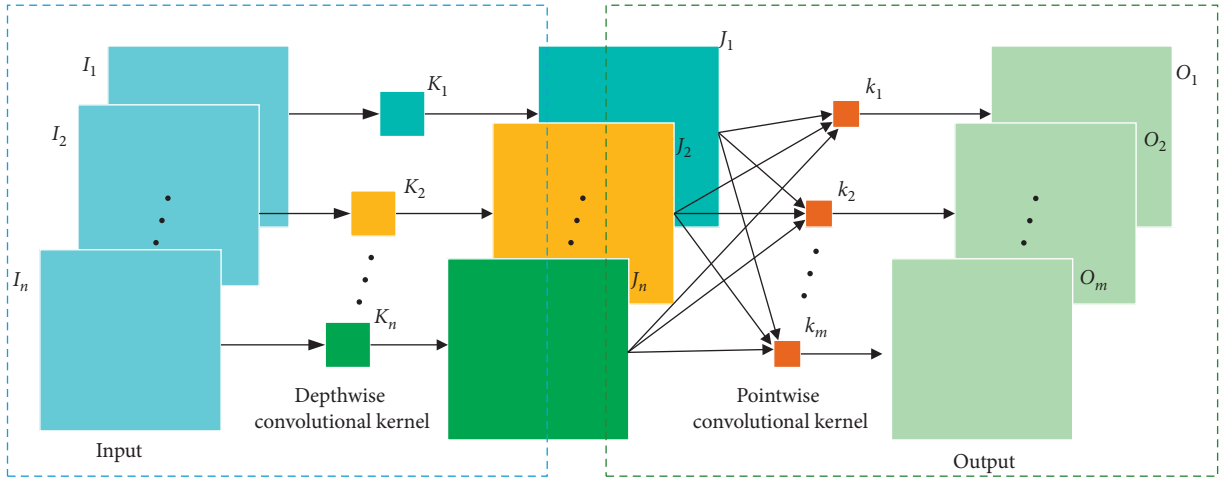


Figure 4: Depthwise separable convolutions (DSConv).

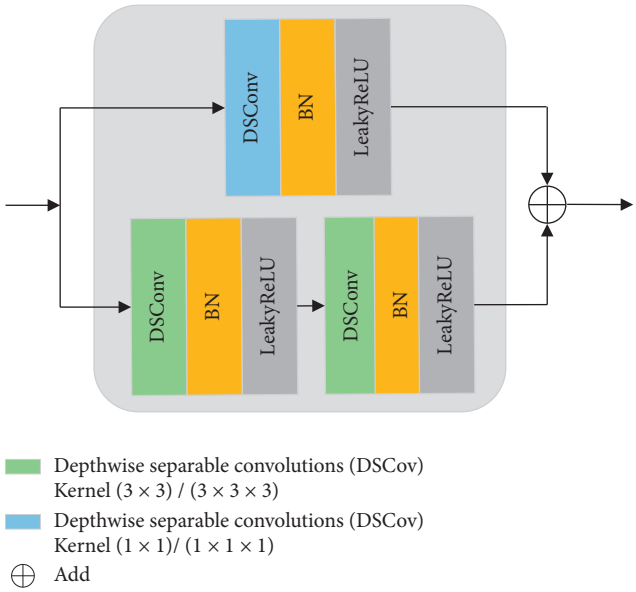


Figure 5: The structure of the DSCR module.

global and local features through an additional layer to obtain real Rician distribution features. Finally, we use the REC module to obtain a predicted clean MR image $\widehat{X}$.

DCRNet. The proposed DCRNet framework is a cascade of 18 DCR modules with different R. The kernel size is $3 \times 3/3 \times 3 \times 3$ for 2D slices/3D patches. Dilated convolution with a large R behaves well for low-frequency noise removal. When R is too large, it is difficult to capture some small contextual information, which will cause the waste of receptive fields. If R is 1, it is the same as the traditional convolution in each channel. In DCRNet, to ensure that all feature maps have the same size as the input, we symmetrically pad zeros around the boundaries before applying the convolution operation. As the convolutional layer increases, the range of the receptive field will gradually increase. In addition, a gridding problem is known to exist in dilated convolution [47]. To solve these problems, considering the size of the input in our experiments, we applied DCR modules with different dilation rates. Therefore, the dilated rate of each layer is set to 1, 1, 1, 1, 1, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, and 1. The final receptive field is 61. Multiscale global features are extracted by using multiple DCR modules with different dilation rates. Each module has 16 filters. The implementations can avoid the gridding effects and reduce the influence of unrelated information.

DSCRNet. DSCRNet is further used to compensate for the local information ignored by expanding the receptive field. It is a cascade of 18 DSCR modules. The size of the convolution kernel of each module is $3 \times 3/3 \times 3 \times 3$. Each module also has 16 filters.

We fused the features extracted from each module of DCRNet and DSCRNet to gradually realize the complementarity of global and local information. This process particularly helps to preserve critical image features in global regions and local regions. Therefore, the proposed 3D-Parallel-RicianNet model will have better denoising ability than other methods.

REC Module. After a convolution layer, we obtain the estimated deviation $f(Y;\Theta)$ and then use $\widehat{X} = Y - f(Y;\Theta)$ to obtain a predicted clean MR image.

2.5. Loss Function. Our loss function uses the mean squared error (MSE) as follows:

$$\iota(\Theta) = \frac{1}{N}\sum_{i=1}^{N} Y_i - f(Y_i;\Theta) - X_i^2. \tag{3}$$

where X_i is the i^{th} noise-free image, Y_i is the corresponding noisy image, and Θ denotes the network parameters. We minimize this loss function to learn the output noise-free image $(Y_i - f(Y_i;\Theta))$.

3. Experiments Results and Analysis

3.1. Dataset Description. To validate the performance of the proposed 3D-Parallel-RicianNet, extensive experiments

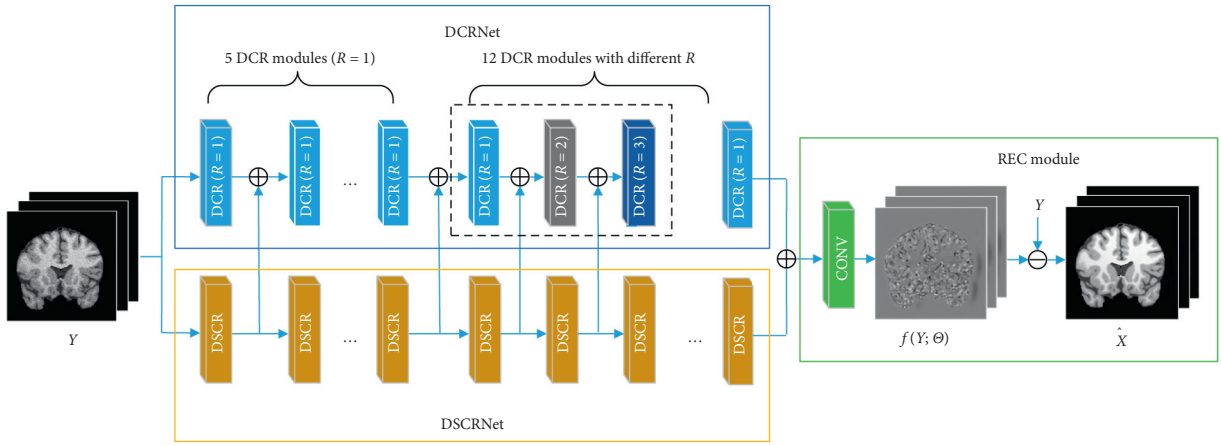


Figure 6: The proposed network architecture for MR image denoising.

were performed on both public simulated and clinical datasets.

For simulated experiments, the BrainWeb dataset [48, 49] was used. In this work, we obtained 18 T1-weighted (T1w) MR images with different noise levels (1%, 3%, 5%, 7%, and 9%). The size of the image is $181 \times 217 \times 181$, and its resolution is $1 \times 1 \times 1\text{mm}^3$. The brain skull is stripped by the skull mask. To further speed up the training process and obtain fewer redundant areas, we cropped the edges of the image, and the image size is $160 \times 192 \times 160$.

One critical problem of the deep learning approach is weak generalization applicability. Networks trained on one dataset from a specific manufacturer or setting may not perform well for a different dataset. The noise in the simulation dataset is assumed to come from single coil acquisition systems. However, clinical MR image noise distributions come from multiple coils, and these noises are subject to a noncentral Chi distribution with a sum-of-squares (SoS) reconstruction. Actually, the Rician distribution is a special case of the noncentral Chi distribution [50] and varies spatially in real MR images [37].

To verify the generalization ability of the proposed model, we carried out experiments on real datasets. For the first clinical experiment, the well-known IXI dataset [51] was used, which was collected from 3 different hospitals. We randomly selected 100 T1w brain images from the Hammersmith dataset. The image size is $256 \times 256 \times 150$, and the voxel resolution is $0.9375 \times 0.9375 \times 1.2\text{mm}^3$. Sixty images were randomly selected as the training set, 20 images for validation, and the other 20 images for testing. In this dataset, we manually added different levels of Rician noise to simulate the noisy image [26]. The brain skull was stripped by the VolBrain method [52].

For another experiment, we randomly selected 35 T1w images in ADNI [53]. Each of these samples contained $192 \times 192 \times 160$ voxels with $1.2 \times 1.25 \times 1.25\text{mm}^3$ voxel resolution. For the experiment, the original scan was resized to dimensions of $256 \times 256 \times 128$. The brain skull was also stripped by the VolBrain method. Due to the lack of knowledge about the noise level in real data, we used the variance-stabilization approach to estimate the Rician noise level of ADNI data, which was approximately 3% [54]. Hence, we selected IXI models trained with a 3% noise level to test ADNI data.

The last dataset comes from the Combined Healthy Abdominal Organ Segmentation (CHAOS) challenge [55, 56]. The dataset included 40 abdominal T1w MR images. On average, each volume size is $256 \times 256 \times 36$, and the noise level is unknown. We adjusted the image to $256 \times 256 \times 64$ through zero-padding operations to be uniform. To substantiate the robustness and generalization capability of the proposed framework, we employed this dataset for our experiments, splitting it into subsets of 25, 5, and 10 subjects that were used for training, validation, and testing.

3.2. Training Details. We use two strategies for training on the three datasets of BrainWeb, IXI-Hammersmith, and CHAOS: 2D slice-based training and 3D patch-based training. For 2D training, we extracted 2D coronal slices from 3D data in the BrainWeb dataset. We obtained 2880 slices by rotating 90° and mirroring, with 1920 slices for training, 384 slices for validation, and 576 slices for testing. In the IXI-Hammersmith dataset, we cropped the image to $256 \times 256 \times 128$ and tested it in all clinical brain datasets. We extracted 7680 slices for training, 2560 slices for validation, and 2560 slices for testing in the sagittal plane. In the CHAOS dataset, using rotation and mirroring to expand the data, we obtained 6400 sagittal slices for training, 1280 sagittal slices for validation, and 640 sagittal slices for testing.

For patch-based training, 3D data in the BrainWeb dataset was also expanded by rotation and mirroring. To reduce memory burden, we used patches with a size of $64 \times 64 \times 64$ voxels. A sliding window strategy with a stride of $16 \times 32 \times 16$ was then used to obtain 3675 patches to train the 3D model. Using the same strategy as the BrainWeb dataset, 4500 training patches, 1500 validation patches, and 1500 test patches were extracted from IXI-Hammersmith with a step size of $48 \times 48 \times 32$. We used rotating and mirroring to expand the CHAOS dataset before extracting patches and finally obtained 4900 training patches, 980

validation patches, and 490 test patches with a stride of $32 \times 32 \times 64$. In the training stage, since the CHAOS dataset did not have clean images and the noise level was unknown, we used the 5% noise model trained by IXI-Hammersmith to estimate clean images as ground truth. In the testing stage, we applied the trained network to patches of the test set. The resultant predictions were averaged in the overlapping regions.

All training was conducted using a deep learning acceleration computing service, which is configured with a 2.20 GHz Core i7-8750H CPU, an NVIDIA GeForce GTX 1070 (8G) GPU, and 16 GB RAM. All the deep learning models were implemented with the publicly available TensorFlow framework and Keras artificial neural network library. In the training process, the learning rate was set to 1e-3. We used Adam optimization.

3.3. Evaluation Methods. Six kinds of deep learning models were trained: CNN-DMRI [28], RicianNet [29], 2D-DCRNet, 2D-Parallel-RicianNet, 3D-DCRNet, and 3D-Parallel-RicianNet. We compared these six deep learning models with four traditional denoising methods: NLM, BM3D, ODCT3D [57], and PRI-NLM3D [57]. In the NLM method, the fastNLMMeansDenoising function is selected, where the template size is 7×7 and the filter strength is 15.

Three quantitative metrics were employed to evaluate the denoising performance of these methods. The first was the peak signal-to-noise ratio (PSNR). A high PSNR generally denotes good denoising performance. The second was the structural similarity index (SSIM), which measured the structural similarity between the ground-truth and denoised images. The last one was entropy, which reflected the amount of image information. We used the natural logarithm in the entropy metric.

3.4. Simulated Results. The quantitative results of NLM, BM3D, ODCT3D, PRI-NLM3D, CNN-DMRI, 2D-DCRNet, 2D-Parallel-RicianNet, 3D-DCRNet, and 3D-Parallel-RicianNet on T1w images with different noise levels (1%, 3%, 5%, 7%, 9%) are illustrated in Tables 1–3.

Tables 1 and 2 depict the PSNR and SSIM results, respectively. We can observe that the PSNR values of 3D-Parallel-RicianNet are obviously higher than those of the other methods at all noise levels. In Table 2, the SSIM values of 3D-Parallel-RicianNet are closer to 1, which is higher than those of the other methods under all noise levels except PRI-NLM3D at the 7% noise level. This indicates that our proposed model has good denoising performance with good anatomical structure preservation.

Table 3 shows the entropy results of 10 methods. We find that the proposed 3D-Parallel-RicianNet can obtain the lowest entropy under all five noise levels. Hence, considering the three metrics in 3 tables, we find that our method has better noise reduction performance. In addition to visual quality, another important aspect of the MR image denoising method is the time complexity. We give running times for different methods in Table 4. It is clear that 3D-DCRNet and our proposed 3D-Parallel-RicianNet are much faster than other methods. Once the deep learning-based method finishes training, forward propagation is very fast. In Table 5, our method has the fewest parameters, which means that our network does not need too much computational power. From this, we can see that our model has competitive advantages for small data sets.

Figures 7 and 8 provide a visual comparison for T1w images from testing data under 3% and 9% noise levels using 10 methods. The zoomed-in regions of the denoised images are shown to observe noticeable details. In Figure 7, all methods can achieve good performance under low-level noise circumstances. However, traditional methods suffer from obvious oversmoothing effects and distort some important details. Among deep learning methods, the images processed by CNN-DMRI, 2D-DCRNet, 2D-Parallel-RicianNet, and 3D-DCRNet have obvious Rician noise. RicianNet increases the brightness of the brain area and makes it difficult to clearly observe the anatomical structure. Figure 7 shows that the 3D-Parallel-RicianNet denoising method gives better results and preserves the key information in the image.

While the noise level increases, the traditional methods suffer from obvious oversmoothing effects, as shown in Figure 8. CNN-MRI and RicianNet models still have some noise and suffer from slight oversmoothing of textured regions. By using the DCR module, 2D-DCRNet and 3D-DCRNet have a strong denoising ability globally for 2D slice-based and 3D patch-based cases. However, without considering local structural features, the DCRNet model loses some important local details in the denoising process. Hence, by combining global features of DCRNet and local features of DSCRNet, the proposed 3D-Parallel-RicianNet can preserve finer detailed structures in homogeneous areas, and it obtains the most consistent results with noise-free images. Hence, our 3D-Parallel-RicianNet method can better retain the key information in denoised MR images, which is useful for improving the precision of clinician diagnosis.

3.5. Clinical Results

3.5.1. Results from the IXI-Hammersmith Dataset. To validate the performance of the proposed 3D-Parallel-RicianNet, ten denoising methods were compared on different clinical data sets.

Figures 9–11 summarize the three metrics in the IXI-Hammersmith dataset with 10 methods under different noise levels. At a noise level of 1%, the PRI-NLM algorithm achieves denoising performance comparable to that of 3D-Parallel-RicianNet in terms of PSNR. At noise levels above 5%, the proposed model produces higher PSNRs than the competing methods. In particular, in Figure 10, we can see that the 3D-Parallel-RicianNet model consistently yields SSIMs higher than the other nine methods for all noise levels. From the perspective of entropy, our method had a low entropy value. These results indicated that the 3D-Parallel-RicianNet model had a strong denoising ability.

Table 1: PSNRs on different noise levels from BrainWeb dataset with 10 methods.

Methods	1%	3%	5%	7%	9%
NLM	34.1381	32.1259	29.9717	29.4599	27.8319
BM3D	35.3151	33.1937	31.0144	30.4890	28.9093
ODCT3D	48.1295	36.2756	31.5857	30.5124	28.2452
PRI-NLM3D	49.8923	36.8192	31.9597	30.9418	28.5961
CNN-DMRI	47.8125	35.4720	30.8958	29.3533	27.2152
RicianNet	43.4145	37.0095	27.4807	27.6777	28.8616
2D-DCRNet	46.8168	38.6786	34.6277	32.4664	30.5232
2D-Parallel-RicianNet	50.9072	41.8099	38.8218	35.9767	34.5207
3D-DCRNet	48.5120	39.1336	35.1466	32.7280	31.0916
3D-Parallel-RicianNet	**51.7192**	**43.9950**	**40.9218**	**37.6896**	**37.1069**

Table 2: SSIMs on different noise levels from BrainWeb dataset with 10 methods.

Methods	1%	3%	5%	7%	9%
NLM	0.9669	0.9605	0.9539	0.9521	0.9454
BM3D	0.9768	0.9725	0.9673	0.9649	0.9584
ODCT3D	0.9992	0.9959	0.9903	0.9857	0.9778
PRI-NLM3D	0.9995	0.9967	0.9922	**0.9889**	0.9823
CNN-DMRI	0.9987	0.9891	0.9727	0.9533	0.9312
RicianNet	0.9606	0.8906	0.9221	0.7246	0.6319
2D-DCRNet	0.9986	0.9900	0.9906	0.9535	0.9346
2D-Parallel-RicianNet	0.9992	0.9933	0.9864	0.9685	0.9722
3D-DCRNet	0.9985	0.9898	0.9587	0.9561	0.9373
3D-Parallel-RicianNet	**0.9995**	**0.9982**	**0.9942**	0.9883	**0.9859**

Table 3: Entropy on different noise levels from BrainWeb dataset with 10 methods.

Methods	1%	3%	5%	7%	9%
Noisy image	2.4787	2.5067	2.5266	2.5482	2.5556
NLM	2.4721	2.4581	2.4474	2.4470	2.4324
BM3D	2.4532	2.4476	2.4849	2.4676	2.4803
ODCT3D	2.4516	2.4844	2.4988	2.5102	2.5033
PRI-NLM3D	2.4447	2.4415	2.4291	2.4330	2.4201
CNN-DMRI	2.4499	2.4869	2.5086	2.5313	2.5406
RicianNet	2.4629	2.4715	2.4519	2.4511	2.4606
2D-DCRNet	2.4624	2.4701	2.4742	2.3481	2.4142
2D-Parallel-RicianNet	2.4693	2.3995	2.4073	2.4631	2.1374
3D-DCRNet	2.4556	2.4488	2.4331	2.3340	2.1787
3D-Parallel-RicianNet	**2.4364**	**2.3469**	**2.3275**	**2.2890**	**2.0642**

Figure 12 shows an example of denoising results using 10 methods on the IXI-Hammersmith dataset with 3% noise. It can be seen in the figure that the proposed 3D-Parallel-RicianNet model gives the best denoising results and the denoised image is virtually identical to the ground-truth image. After visual inspection, it can be deduced that the outcome of our proposed 3D-Parallel-RicianNet is improved compared to the others in terms of fine-structure retention and edges.

3.5.2. Results from the IXI-Guys Dataset. Figures 13–15 summarize the PSNR, SSIM, and entropy values using 10 methods on the IXI-Guys dataset. We test the trained model with the IXI-Hammersmith dataset on this IXI-Guys dataset, which reflects network generalization on other nontrained datasets. The 3D-Parallel-RicianNet shows the most robust performance among the tested methods in terms of PSNR, SSIM, and entropy. In particular, our model still achieves better denoising ability than other methods at higher noise levels.

Figure 16 shows an example of denoising results obtained with 10 methods on data from the IXI-Guys dataset at the 3% noise level. Consistent with the denoising performance on the IXI-Hammersmith dataset, the proposed 3D-Parallel-RicianNet method provided the best denoising result and removed the image noise more robustly than the other methods on the IXI-Guys dataset. Particularly in the region indicated by the red line, the 3D-Parallel-RicianNet model achieved better visual results.

Table 4: Execution time on different noise levels from BrainWeb dataset with 10 methods.

Methods	1%	3%	5%	7%	9%	Average
NLM	10.55	10.54	10.58	10.54	10.55	10.55
BM3D	28.26	28.13	27.29	27.37	28.25	27.86
ODCT3D	23.54	17.56	18.02	14.75	14.54	17.68
PRI-NLM3D	14.07	12.50	12.92	13.59	13.28	13.27
CNN-DMRI	1.13	1.11	1.13	1.12	1.12	1.12
RicianNet	1.71	1.65	1.65	1.69	1.66	1.67
2D-DCRNet	1.27	1.32	1.34	1.33	1.29	1.31
2D-Parallel-RicianNet	1.18	1.14	1.13	1.12	1.12	1.14
3D-DCRNet	0.94	0.92	0.91	0.91	0.92	0.92
3D-Parallel-RicianNet	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**

Table 5: Number of parameters of different networks.

Method	CNN-DMRI	RicianNet	3D-Parallel-RicianNet
Number of parameters	1,444,929	5,346,114	395,405

3.5.3. Results from the ADNI Dataset. This subsection is devoted to verifying the consistency of the proposed approach on the ADNI dataset. Because noise-free images are unavailable, entropy is measured and used as the quantitative metric. The results are shown in Figures 17 and 18.

As shown in Figure 17, although the RicianNet and 2D-DCRNet remove noise, they suffer from obvious over-smoothing effects, and it is difficult to identify the key anatomical structures. In addition, the denoising effect is not satisfactory when using BM3D, ODCT3D, PRI-NLM3D, CNN-DMRI, 2D-Parallel-RicianNet, and 3D-DCRNet. The results of these methods still contain substantial noise and miss some of the structural details. It can be noted that 3D-Parallel-RicianNet retains the details better than other methods.

According to Figure 18, the entropy results of denoised MR images in the ADNI dataset using different processing methods are compared. We find that 3D-Parallel-RicianNet

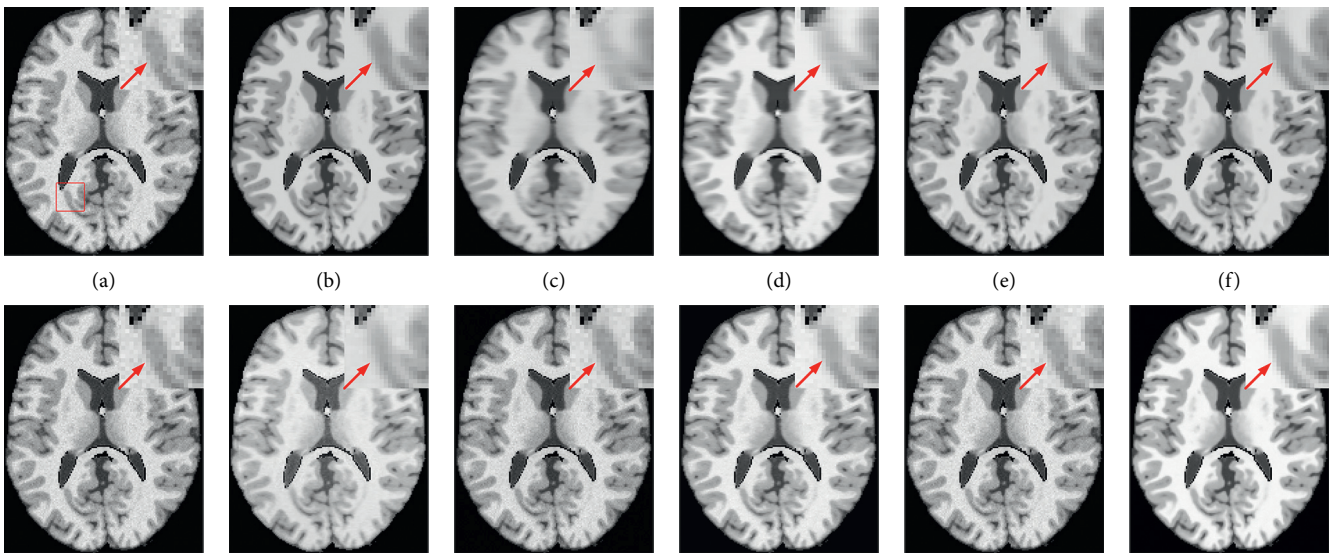


Figure 7: Denoising effect of different methods with 3% noise level. (a) Noisy image; (b) noise-free image; (c) NLM; (d) BM3D; (e) ODCT3D; (f) PRI-NLM3D; (g) CNN-DMRI; (h) RicianNet; (i) 2D-DCRNet; (j) 2D-Parallel-RicianNet; (k) 3D-DCRNet; (l) 3D-Parallel-RicianNet.

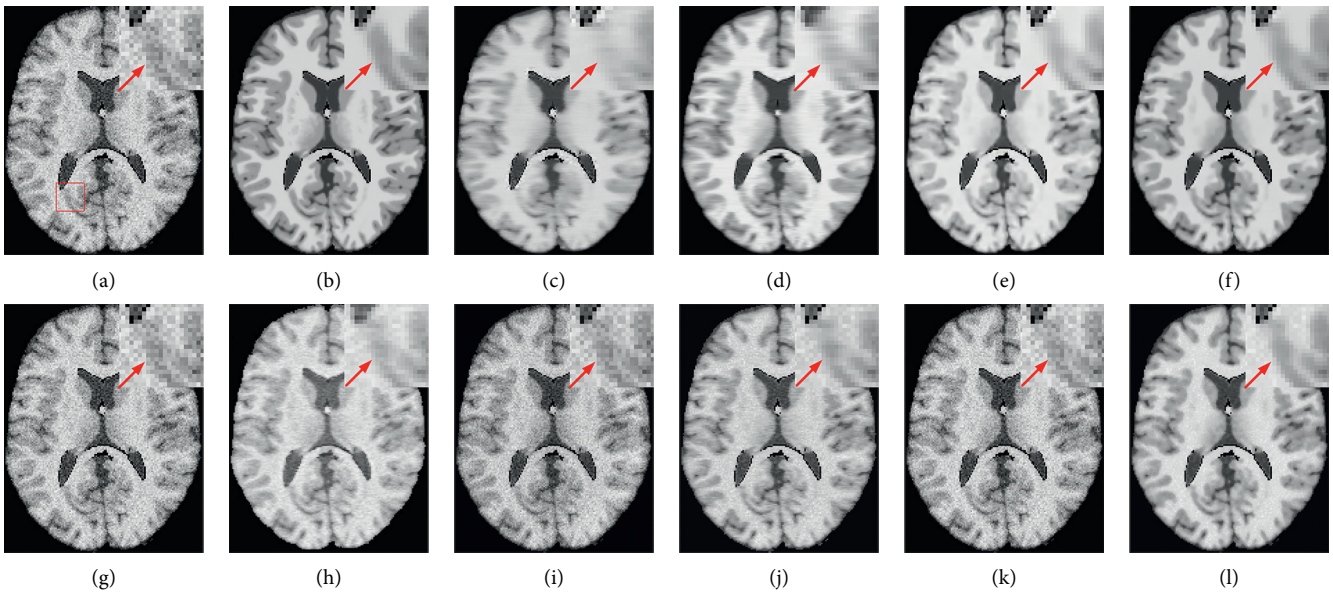


Figure 8: Denoising effect of different methods with 9% noise level. (a) Noisy image; (b) noise-free image; (c) NLM; (d) BM3D; (e) ODCT3D; (f) PRI-NLM3D; (g) CNN-DMRI; (h) RicianNet; (i) 2D-DCRNet; (j) 2D-Parallel-RicianNet; (k) 3D-DCRNet; (l) 3D-Parallel-RicianNet.

achieves the lowest entropy value. Combined with Figure 17, we find that our method not only effectively removes noise but also preserves more useful key information in images. Hence, our 3D-Parallel-RicianNet method has strong generalization ability and strong robustness. These experimental results once again demonstrate the advantages of our proposed model.

3.5.4. Denoising of Real Abdominal MR Data. In this subsection, we performed denoising for abdominal MR images by the proposed network. We compared three denoising methods, and the experimental results are shown in Table 6.

Table 6 shows that the PSNR of our method can reach 39.7090, which is higher than those of BM3D, CNN-DMRI, and RicianNet. On SSIM, RicianNet is lower than BM3D and

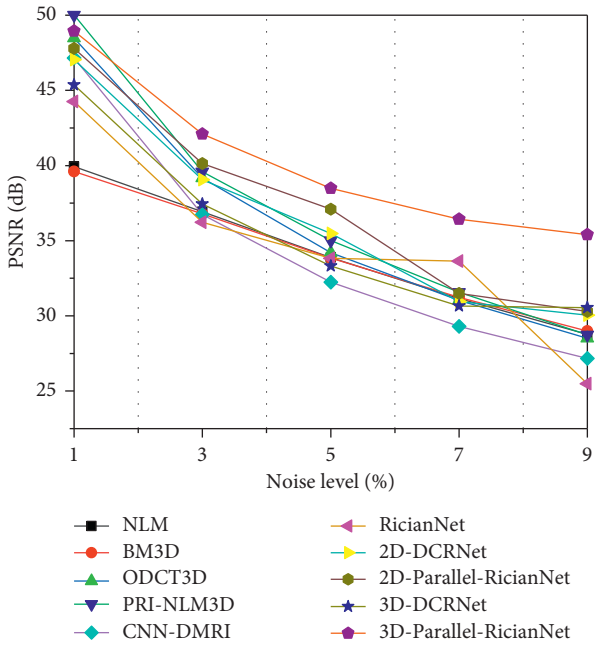


Figure 9: PSNRs using 10 methods under different noise levels for the IXI-Hammersmith dataset.

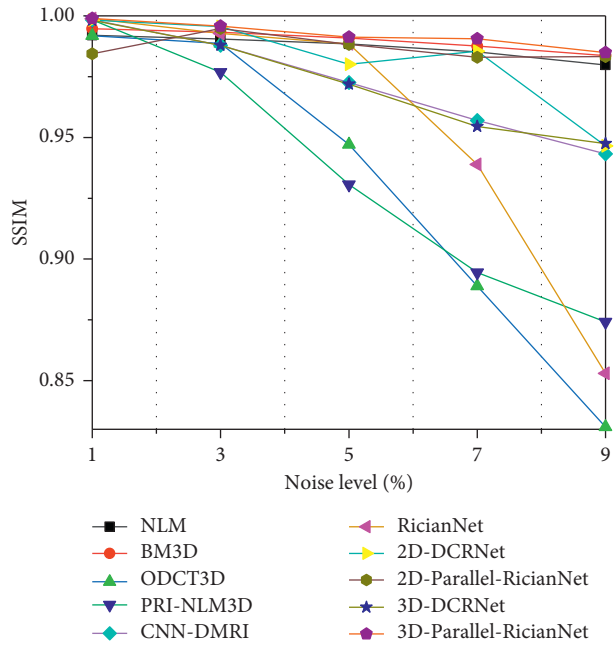


Figure 10: SSIMs using 10 methods under different noise levels for the IXI-Hammersmith dataset.

CNN-DMRI, indicating that although RicianNet can remove noise, it cannot retain the structure information of the image. Our method can still obtain the highest SSIM value. We show the denoising results of the four methods in Figure 19. It can be seen from the figure that our method can not only remove noise but also preserve the key anatomical position information in the image completely.

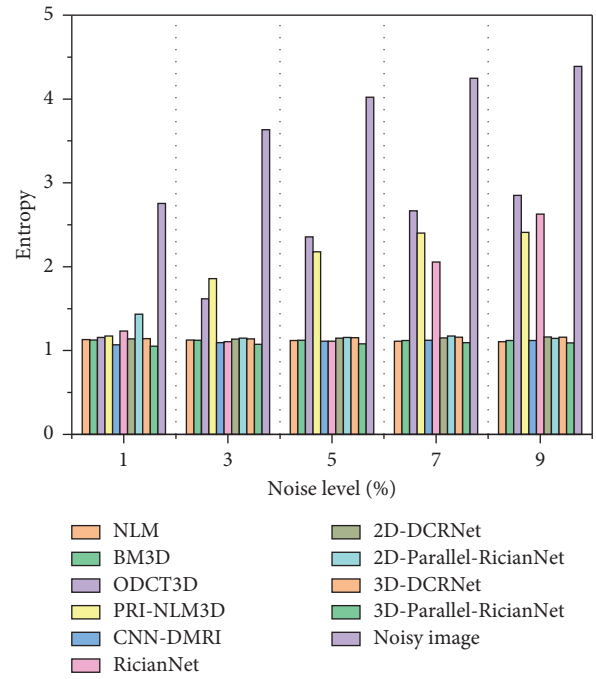


Figure 11: Entropy using 10 methods under different noise levels for the IXI-Hammersmith dataset.

3.6. Comparisons of the Results with Different Spatial Resolutions. The image resolution affects the quality of the image. Generally, when the image resolution is smaller, the denoising ability of the model is significantly reduced. In this part, we use the BrainWeb dataset to verify the denoising effect of images with different resolutions at a noise level of 3%. The results are shown in Table 7.

From Table 7, it can be observed that the proposed 3D-Parallel-RicianNet outperforms other methods tested among the different spatial resolutions. For BM3D and RicianNet, they are difficult to remove noise at low spatial resolutions. It is noted that noise cleaning appears to have a consistent effect when different spatial resolutions are relatively close, such as $0.9375 \times 0.9375 \times 0.9375\text{mm}^3$ and $1 \times 1 \times 1\text{mm}^3$. However, it should be noted that some loss of contrast and spatial resolution is possible. Once the difference between the resolutions becomes larger, the denoising effect will also change significantly, such as $1 \times 1 \times 1\text{mm}^3$ and $2 \times 2 \times 2\text{mm}^3$. In addition, the PSNRs of deep learning methods decrease significantly with decreasing spatial resolution, while the SSIM values are relatively close, indicating that deep learning methods recover most of the complex anatomical structures. Compared to other methods, our model has a more balanced denoising ability at different spatial resolutions, and the mean value of PSNR can reach 41.6373. Since the proposed 3D-Parallel-RicianNet can extract the global and local features in the noisy image and restore the clean image, it can still maintain denoising ability at low spatial resolution.

3.7. Comparisons of the Results with Different Brain Tissues. Based on MR imaging technique, key brain tissues like gray matter (GM), white matter (WM), and cerebrospinal

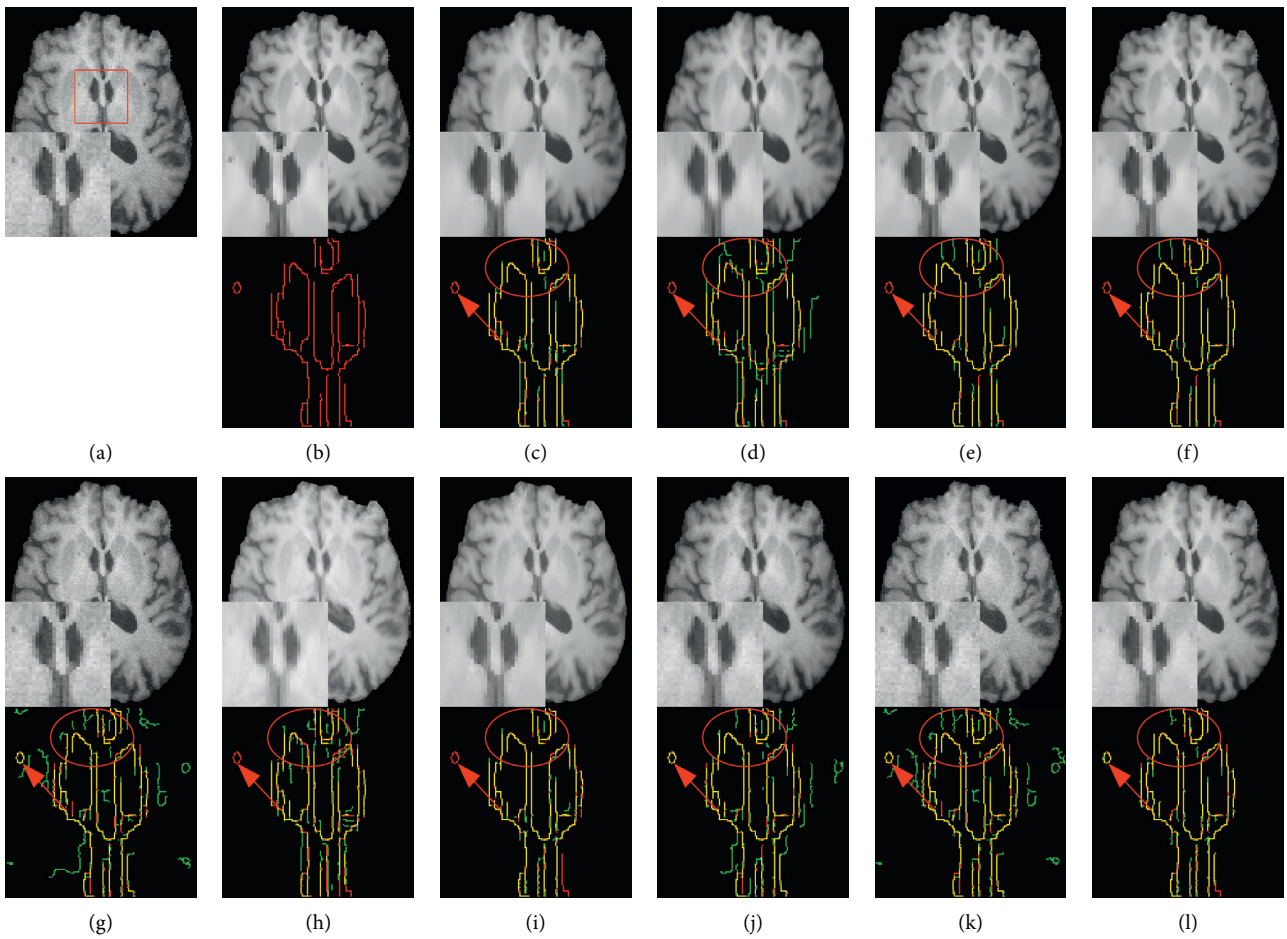


Figure 12: The denoising effect for IXI-Hammersmith dataset with different methods at 3% noise level. (a) Noisy image; (b) noise-free image; (c) NLM; (d) BM3D; (e) ODCT3D; (f) PRI-NLM3D; (g) CNN-DMRI; (h) RicianNet; (i) 2D-DCRNet; (j) 2D-Parallel-RicianNet; (k) 3D-DCRNet; (l) 3D-Parallel-RicianNet. Each method below shows the corresponding edge detection image of the enlarged area (green), and the yellow represents the overlapping area with the noise-free image (red).

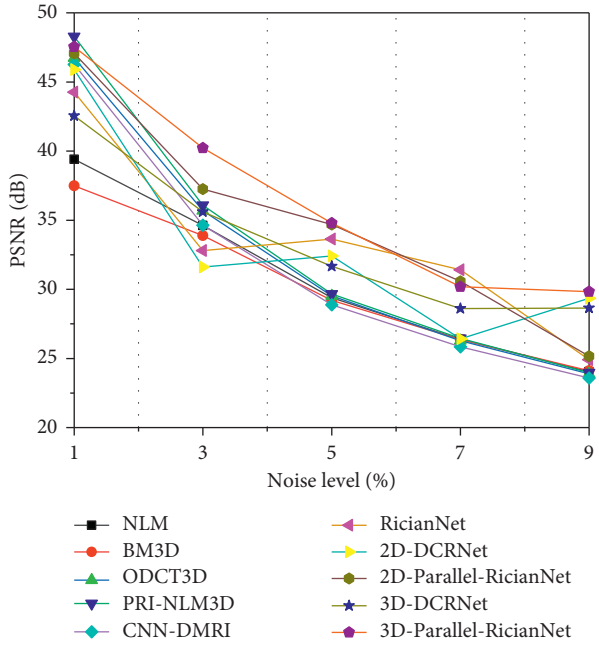


Figure 13: PSNRs using 10 methods under different noise levels for the IXI-Guys dataset.

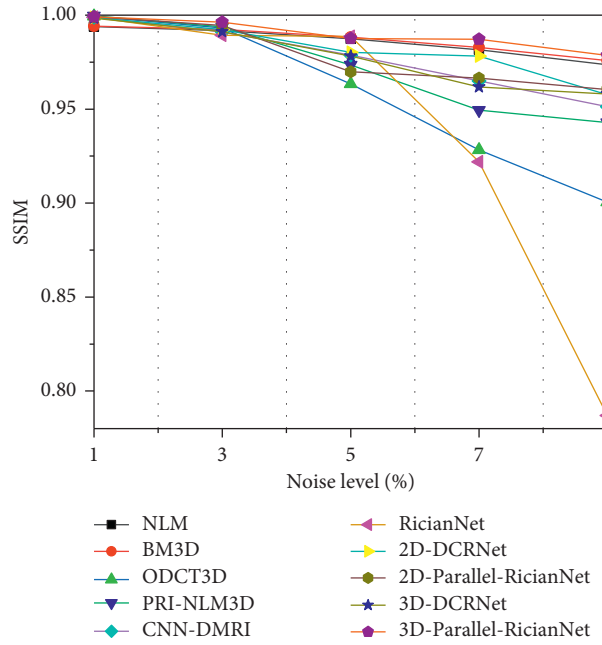


Figure 14: SSIMs using 10 methods under different noise levels for the IXI-Guys dataset.

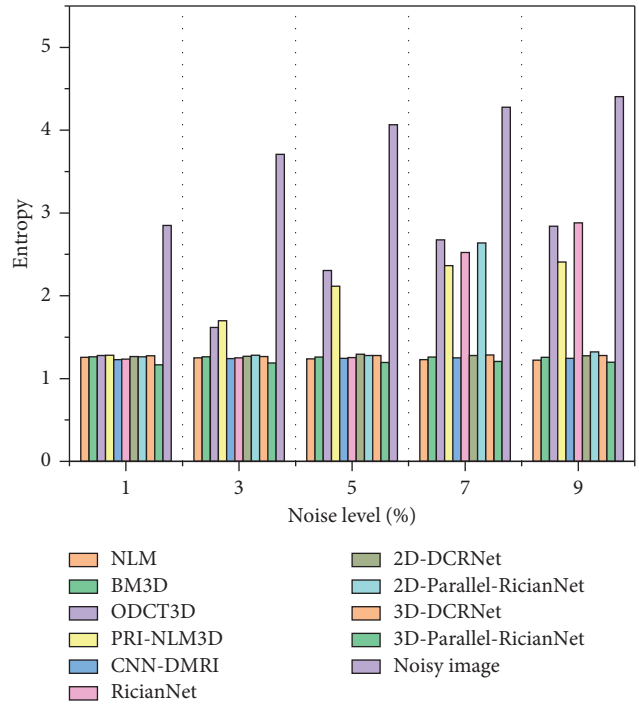


Figure 15: Entropy using 10 methods under different noise levels for the IXI-Guys dataset.

fluid (CSF) become visible. These three tissues help visualize brain structures and guide surgery but noise can affect the interpretation of brain tissue [58]. To evaluate the denoising effectiveness of 3D-Parallel-RicianNet on different brain tissues, state-of-the-art methods BM3D, CNN-DMRI, and RicianNet are compared in Table 8. The proposed model can achieve better PSNR and SSIM results than the competing methods in different brain tissues. In particular, in CSF, we can see that the PSNR of 3D-Parallel-RicianNet can reach 51.9105. We also show the different brain tissue denoising results of four denoising methods in Figure 20. These experimental results once again demonstrate the advantages of the proposed model.

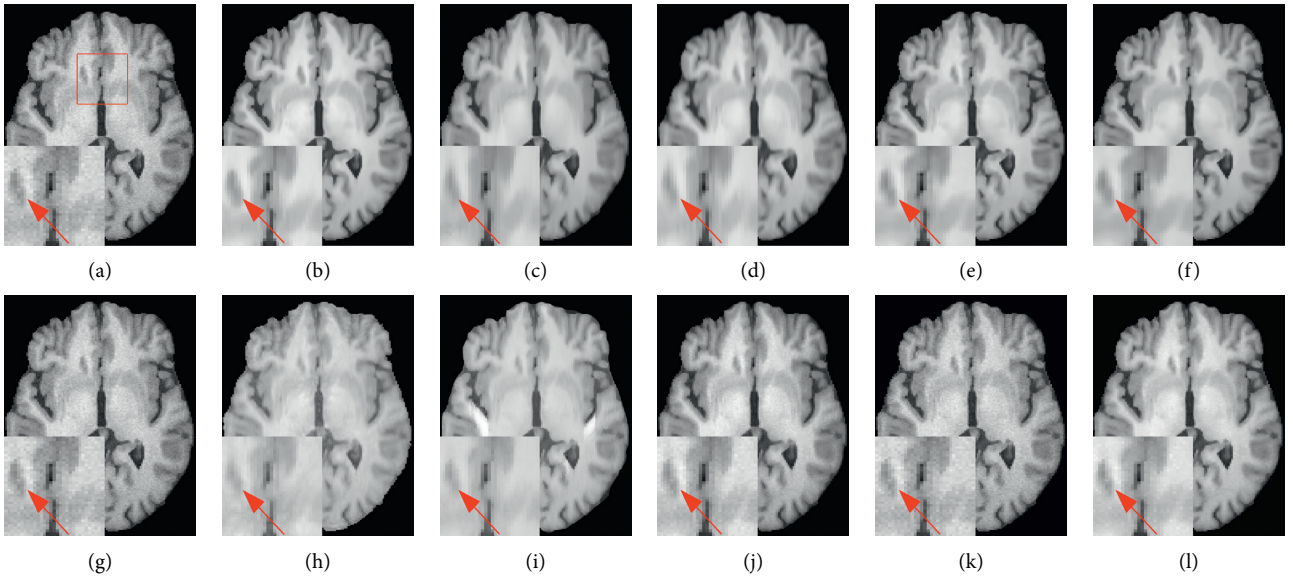


Figure 16: The denoising effect for IXI-Guys dataset with different methods at 3% noise level. (a) Noisy image; (b) noise-free image; (c) NLM; (d) BM3D; (e) ODCT3D; (f) PRI-NLM3D; (g) CNN-DMRI; (h) RicianNet; (i) 2D-DCRNet; (j) 2D-Parallel-RicianNet; (k) 3D-DCRNet; (l) 3D-Parallel-RicianNet.

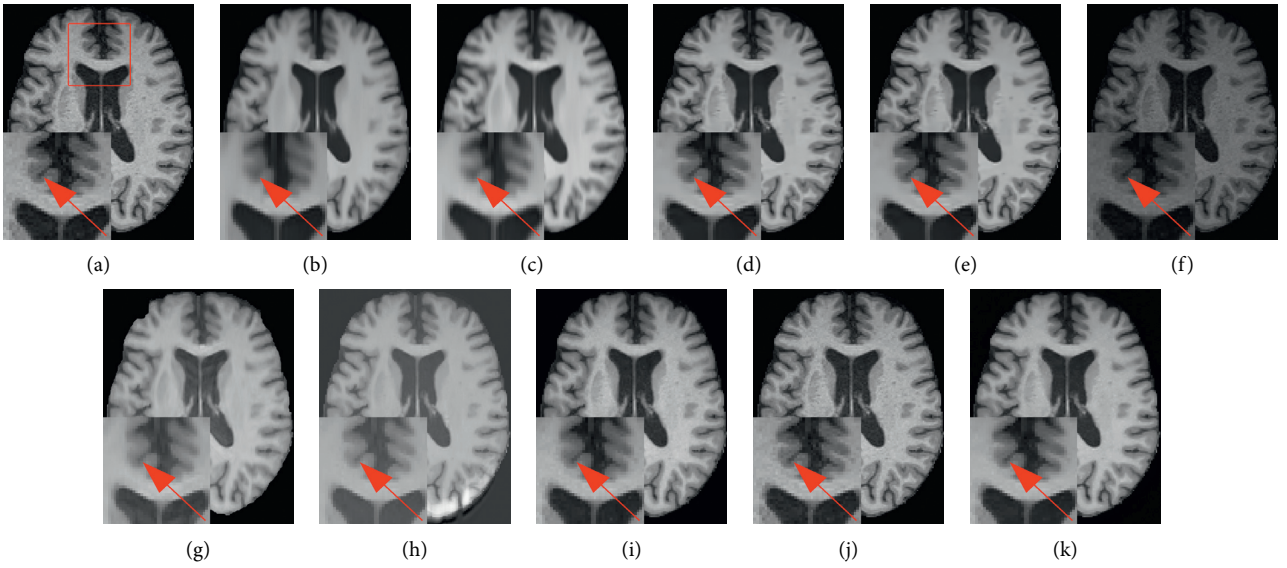


Figure 17: The denoising effect of ADNI set with different methods at 3% noise level. (a) Noisy image; (b) NLM; (c) BM3D; (d) ODCT3D; (e) PRI-NLM3D; (f) CNN-DMRI; (g) RicianNet; (h) 2D-DCRNet; (i) 2D-Parallel-RicianNet; (j) 3D-DCRNet; (k) 3D-Parallel-RicianNet.

3.8. Variants of the **R** *Setting in the DCR Module.* In our model, the DCR module of different R is our key component. The PSNR and SSIM are recorded in Table 9 by different R settings at the 3% noise level in the BrainWeb dataset. We conducted three experiments, each using the same dilation rate for the 18 DCR modules. The final receptive fields are 37, 73, and 109. Combining Tables 1 and 2 and Table 8, we can find that our hybrid dilation rate can reach the highest PSNR and SSIM. When $R = 2$ and 3, the receptive field has already caused waste. In addition, using the same dilation rates can easily cause gridding effects. There is a lack of correlation between the feature maps extracted in this way, and an accurately predicted result cannot be obtained in the end. Therefore, our model can achieve superior denoising performance.

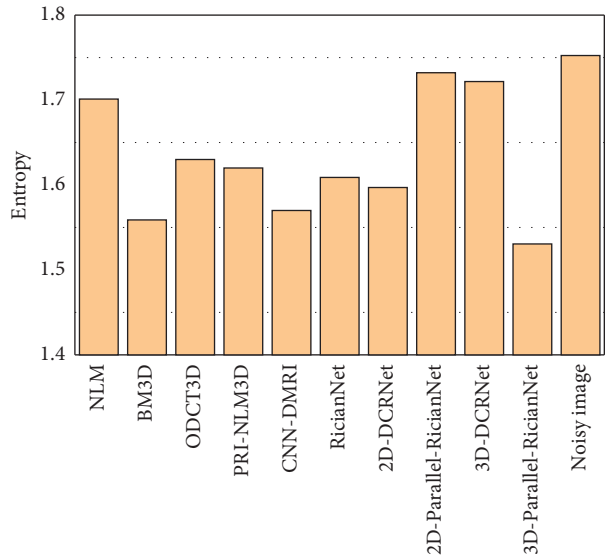


Figure 18: The entropy of the denoised images from the ADNI database.

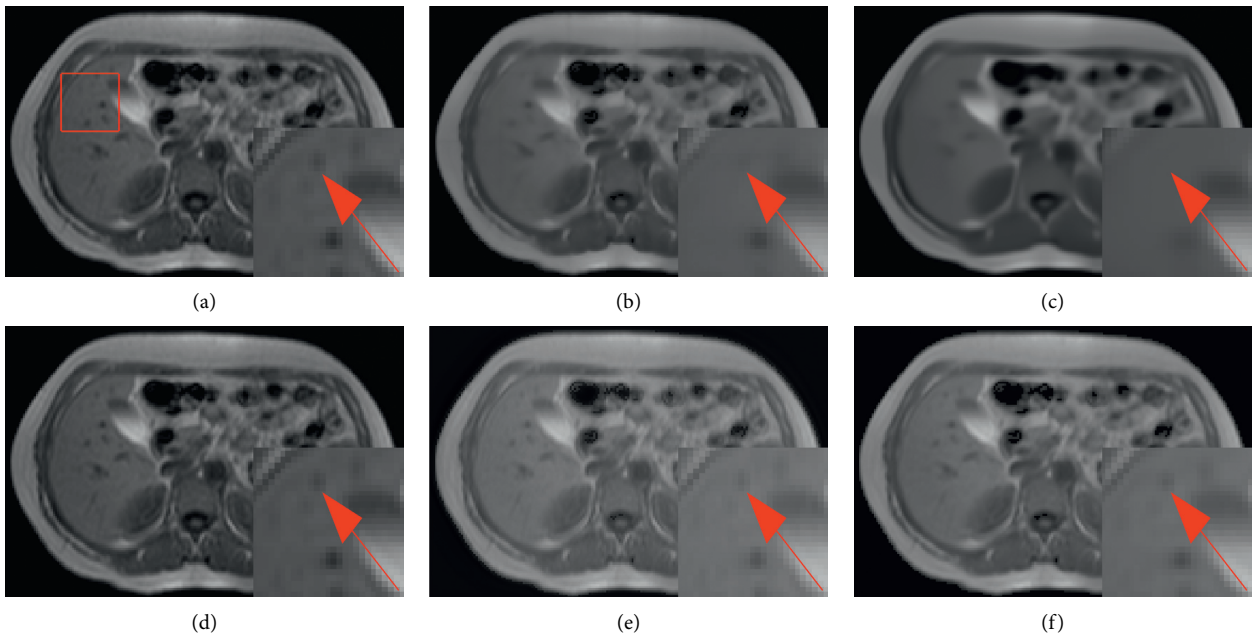


Figure 19: The denoising effect of CHAOS set with different methods at 3% noise level. (a) Noisy image; (b) noise-free image; (c) BM3D; (d) CNN-DMRI; (e) RicianNet; (f) 3D-Parallel-RicianNet.

Table 6: The PSNR and SSIM on different noise levels from CHAOS dataset with 4 methods.

Method	BM3D	CNN-DMRI	RicianNet	3D-Parallel-RicianNet
PSNR	31.9167	32.0655	35.2577	**39.7090**
SSIM	0.9862	0.9867	0.9258	**0.9941**

Table 7: PSNR (top) and SSIM (bottom) comparisons of different algorithms with different spatial resolutions.

Spatial resolutions (mm^3)	Method			
	BM3D	CNN-DMRI	RicianNet	3D-Parallel-RicianNet
$0.9 \times 0.9 \times 0.9$	31.8908 0.9709	34.7330 0.9890	35.7466 0.9233	41.5169 0.9970
$0.9375 \times 0.9375 \times 0.9375$	32.1794 0.9718	35.2350 0.9903	36.4030 0.9117	41.9756 0.9973
$1 \times 1 \times 1$	33.1937 0.9725	35.4720 0.9891	37.0095 0.8906	43.9950 0.9982
$1.1 \times 1.1 \times 1.1$	31.7986 0.9677	35.6006 0.9917	36.6172 0.9078	42.2193 0.9976
$1.25 \times 1.25 \times 1.25$	31.3591 0.9633	34.0402 0.9917	36.0542 0.9070	41.9259 0.9974
$1.5 \times 1.5 \times 1.5$	29.8656 0.9481	35.1568 0.9918	31.4012 0.9180	40.3297 0.9961
$2 \times 2 \times 2$	28.2362 0.9230	34.5695 0.9900	23.4731 0.9038	38.4990 0.9933

Table 8: PSNR and SSIM comparisons of different methods in different brain tissues.

Method	CSF		GM		WM	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BM3D	27.0765	0.9213	19.1709	0.9114	18.2741	0.9350
CNN-DMRI	45.8322	0.9981	39.1790	0.9976	38.7420	0.9973
RicianNet	44.8454	0.9982	41.9853	0.9988	43.6082	0.9986
3D-Parallel-RicianNet	**51.9105**	**0.9996**	**47.5288**	**0.9996**	**48.6238**	**0.9998**

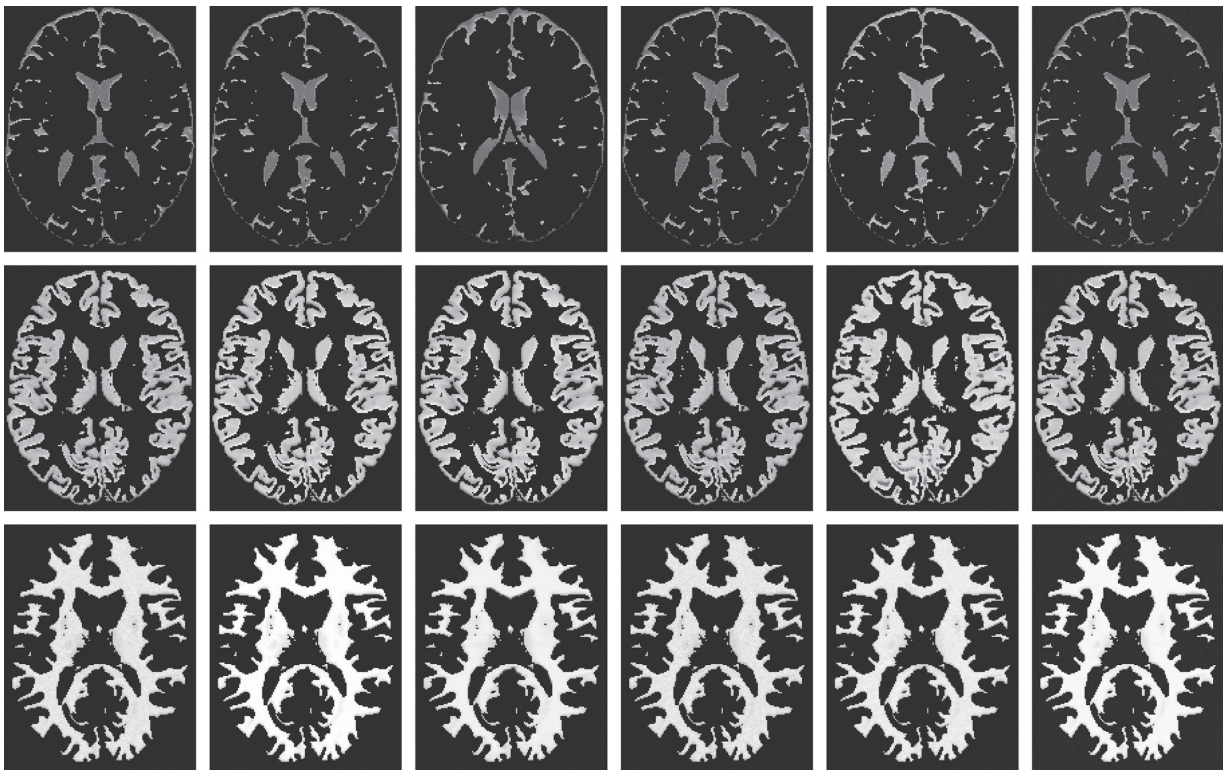


Figure 20: The denoising effect of brain tissues with different methods at 3% noise level. Noisy image (1st column); noise-free image (2nd column); BM3D (3rd column); CNN-DMRI (4th column); RicianNet (5th column); 3D-Parallel-RicianNet (6th column).

Table 9: PSNR and SSIM comparisons in different R settings.

	$R = 1$	$R = 2$	$R = 3$
PSNR	41.0253	40.4788	39.3627
SSIM	0.9956	0.9661	0.9946

4. Discussion and Conclusions

In this work, we propose a parallel denoising residual network based on cascaded DCR and DSCR modules to address the random noise in MR images. The global and local features are extracted by the designed DCRNet and DSCRNet, and then these features are fused together. Hence, global and local information is captured to drive the denoising progress of brain MR images by supervised network learning.

The PSNR, SSIM, and entropy are calculated to compare the proposed method with many existing methods, and the denoising effect of the proposed method is verified on the

BrainWeb simulation data under different noise levels. To test the practicability of the proposed network, experiments on real clinical MR images show that the proposed method is superior to other methods for the IXI-Hammersmith, IXI-Guys, ADNI, and CHAOS datasets.

In this work, one of our limitations is that although structural information can be retained at high noise levels, there is still a small amount of local noise, as shown in Figure 8. Next, we will continue to study to find a balance between noise removal and structure maintenance at different noise levels. Another critical limitation of our method is the requirement for high-quality noise-free ground-truth images, which are difficult to obtain in real applications. Incorporation of prior knowledge about organ shape and location is key to improving the performance of image analysis approaches. However, in most recently developed medical image analysis techniques, it is not obvious how to incorporate such prior knowledge [59]. Oktay et al. incorporated anatomical prior knowledge into a deep learning method through a new regularization model, and this method showed that the approach can be easily adapted to different medical image analysis tasks (e.g., image enhancement and segmentation) [59]. Furthermore, in [60], the author used morphological component analysis (MCA) to decompose noisy images into cartoon, texture, and residual parts that were considered noise components. Therefore, to circumvent the limitations of our method, we will verify it using multimodality images and incorporate other meaningful priors, such as residual parts, organ shape, and location to mitigate semisupervised denoising tasks in the future.

In conclusion, the results obtained in this paper are encouraging and efficiently demonstrate the potential of our 3D-Parallel-RicianNet method for MR image denoising. This method can not only effectively remove noise in MR images but also preserve enough detailed structural information, which can help to provide high-quality MR images for clinical diagnosis.

Data Availability

The BrainWeb, IXI, ADNI, and CHAOS datasets are publicly available (BrainWeb, https://brainweb.bic.mni.mcgill.ca/brainweb/; IXI, http://brain-development.org/ixi-dataset/; ADNI, http://adni.loni.usc.edu/; and CHAOS, https://chaos.grand-challenge.org/).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). This research was supported in part by the National Natural Science Foundation of China under Grant 61771230, the Shandong Provincial Natural Science Foundation under Grant ZR2016FM40, the Shandong Provincial Jinan Science and Technology Project under Grant (201816082 and 201817001), and the Youth Program of Shandong Provincial Natural Science Foundation under Grant ZR2020QF011.

References

[1] Z. Zhang, D. Xia, X. Han et al., "Impact of image constraints and object structures on optimization-based reconstruction," in *Proceedings of the 4nd International Conference on Image Formation in X-Ray Computed Tomography*, pp. 487–490, Bamberg, Germany, January 2016.

[2] J. Mohan, V. Krishnaveni, and Y. Guo, "A survey on the magnetic resonance image denoising methods," *Biomedical Signal Processing and Control*, vol. 9, pp. 56–69, 2014.

[3] J. D. Kumar and V. Mohan, "Edge detection in the medical MR brain image based on fuzzy logic technique," in *Proceedings of the International Conference on Information Communication and Embedded Systems (ICICES2014)*, pp. 1–9, Chennai, India, February 2014.

[4] K. Gupta and S. K. Gupta, "Image denoising techniques: a review paper," *International Journal of Innovative Technology & Exploring Engineering*, vol. 2, no. 4, pp. 6–9, 2013.

[5] H. M. Ali, "MRI medical image denoising by fundamental filters," in *High-Resolution Neuroimaging-Basic Physical Principles and Clinical Applications*, A. M. Halefoğlu, Ed., pp. 111–124, InTech, Zagreb, Croatia, 2018.

[6] K. N. Chaudhury and S. D. Dabhade, "Fast and provably accurate bilateral filtering," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2519–2528, 2016.

[7] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[8] X. Yang and B. Fei, "A wavelet multi-scale denoising algorithm for magnetic resonance (MR) images," *Measurement Science and Technology*, vol. 22, no. 2, Article ID 025803, 2011.

[9] A. Phophalia, A. Rajwade, and S. K. Mitra, "Rough set based image denoising for brain MR images," *Signal Processing*, vol. 103, pp. 24–35, 2014.

[10] S. P. Awate and R. T. Whitaker, "Feature-preserving MRI denoising: a nonparametric empirical Bayes approach," *IEEE Transactions on Medical Imaging*, vol. 26, no. 9, pp. 1242–1255, 2007.

[11] S. Satheesh and K. V. S. V. R. Prasad, "Medical image denoising using adaptive threshold based on contourlet transform," 2011, https://arxiv.org/abs/1103.4907.

[12] X. Zhang, Z. Xu, N. Jia et al., "Denoising of 3D magnetic resonance images by using higher-order singular value decomposition," *Medical Image Analysis*, vol. 19, no. 1, pp. 75–86, 2015.

[13] N. Leal, E. Zurek, E. Leal, and Esmeide, "Non-Local SVD denoising of MRI based on sparse representations," *Sensors*, vol. 20, no. 5, p. 1536, 2020.

[14] H. V. Bhujle and B. H. Vadavadagi, "NLM based magnetic resonance image denoising - a review," *Biomedical Signal Processing and Control*, vol. 47, pp. 252–261, 2019.

[15] J. Hu, Y. Pu, X. Wu, Y. Zhang, and J. Zhou, "Improved DCT-based nonlocal means filter for MR images denoising," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 232685, 14 pages, 2012.

[16] X. Zhang, G. Hou, J. Ma et al., "Denoising MR images using non-local means filter with combined patch and pixel similarity," *PLoS One*, vol. 9, no. 6, Article ID e100240, 2014.

[17] A. Gautam and M. M. Mathur, "Implementation of NLM and PNLM for de-noising of MRI images," *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 11, pp. 31–37, 2019.

[18] B. Kanoun, M. Ambrosanio, F. Baselice, G. Ferraioli, V. Pascazio, and L. Gomez, "Anisotropic weighted KS-NLM filter for noise reduction in MRI," *IEEE Access*, vol. 8, pp. 184866–184884, 2020.

[19] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: a deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[20] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[21] V. Cherukuri, T. Guo, S. J. Schiff et al., "Deep MR brain image super-resolution using spatio-structural priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 1368–1383, 2020.

[22] J. V. Manj, n and P. Coupe, o"MRI denoising using deep learning and nonlocal averaging," 2019, https://arxiv.org/abs/1911.04798.

[23] C. Liu, X. Wu, X. Yu et al., "Fusing multiscale information in convolution network for MR image super-resolution reconstruction," *Biomedical Engineering Online*, vol. 17, no. 1, p. 114, 2018.

[24] C. H. Pham, A. Ducournau, R. Fablet et al., "Brain MRI super-resolution using deep 3D convolutional networks," in *Proceedings of the IEEE International Symposium on Biomedical Imaging IEEE*, pp. 197–200, Melbourne, Australia, April 2017.

[25] D. Jiang, W. Dou, L. Vosters, X. Xu, Y. Sun, and T. Tan, "Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network," *Japanese Journal of Radiology*, vol. 36, no. 9, pp. 566–574, 2018.

[26] M. Ran, J. Hu, Y. Chen et al., "Denoising of 3D magnetic resonance images using a residual encoder-decoder Wasserstein generative adversarial network," *Medical Image Analysis*, vol. 55, pp. 165–180, 2019.

[27] D. Hong, C. Huang, C. Yang et al., "FFA-DMRI: A network based on feature fusion and attention mechanism for brain MRI denoising," *Frontiers in Neuroscience*, vol. 14, Article ID 577937, 2020.

[28] P. C. Tripathi and S. Bag, "CNN-DMRI: a convolutional neural network for denoising of magnetic resonance images," *Pattern Recognition Letters*, vol. 135, pp. 57–63, 2020.

[29] S. Li, J. Zhou, D. Liang, and Q. Liu, "MRI denoising using progressively distribution-based neural network," *Magnetic Resonance Imaging*, vol. 71, pp. 55–68, 2020.

[30] S. Gregory, Y. Gan, H. Cheng et al., "HydraNet: a multi-branch convolutional neural network architecture for MRI denoising," *Medical Imaging 2021: Image Processing*, vol. 11596, 2021.

[31] H. Aetesam and S. K. Maji, "Noise dependent training for deep parallel ensemble denoising in magnetic resonance images," *Biomedical Signal Processing and Control*, vol. 66, Article ID 102405, 2021.

[32] J. Yang, J. Fan, D. Ai, S. Zhou, S. Tang, and Y. Wang, "Brain MR image denoising for Rician noise using pre-smooth non-local means filter," *Biomedical Engineering Online*, vol. 14, no. 1, p. 2, 2015.

[33] L. He and I. R. Greenshields, "A non-local maximum likelihood estimation method for Rician noise reduction in MR images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 165–172, 2008.

[34] T. Kalaiselvi and N. Kalaichelvi, "Investigation on image denoising techniques of magnetic resonance images," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 4, pp. 104–111, 2018.

[35] S. Aja-Fernandez, C. Alberola-Lopez, and C.-F. Westin, "Noise and signal estimation in magnitude MRI and rician distributed images: a LMMSE approach," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1383–1398, 2008.

[36] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magnetic Resonance in Medicine*, vol. 34, no. 6, pp. 910–914, 1995.

[37] R. W. Liu, L. Shi, W. Huang et al., "Generalized total variation-based MRI Rician denoising model with spatially adaptive regularization parameters," *Magnetic Resonance Imaging*, vol. 32, no. 6, pp. 702–720, 2014.

[38] M. Xu, J. Alirezaie, and P. Babyn, "Low-dose CT denoising with dilated residual network," in *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5117–5120, Honolulu, HI, USA, July 2018.

[39] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, https://arxiv.org/abs/1511.07122.

[40] Y. Peng, L. Zhang, S. Liu, X. Wu, Y. Zhang, and X. Wang, "Dilated residual networks with symmetric skip connection for image denoising," *Neurocomputing*, vol. 345, pp. 67–76, 2019.

[41] X. Zhang, W. Yang, Y. Hu et al., "DMCNN: Dual-domain multi-scale convolutional neural network for compression artifacts removal," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 390–394, Athens, Greece, October 2018.

[42] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, Honolulu, HI, USA, July 2017.

[43] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[44] M. Sandler, A. Howard, M. Zhu et al., "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[45] S. Angizi, Z. He, A. S. Rakin et al., "CMP-PIM: an energy-efficient comparator-based processing-in-memory neural network accelerator," in *Proceedings of the 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, pp. 1–6, San Francisco, CA, USA, June 2018.

[46] R. Imamura, T. Itasaka, and M. Okuda, "Zero-Shot Hyperspectral Image Denoising with Separable Image Prior," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1416–1420, Seoul, South Korea, October 2019.

[47] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[48] C. A. Cocosco, V. Kollokian, R. K. S. Kwan et al., "Brainweb: online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, p. 425, 1997.

[49] R. K.-S. Kwan, A. C. Evans, and G. B. Pike, "An extensible MRI simulator for post-processing evaluation," in *Lecture Notes in Computer Science*, pp. 135–140, Springer, Berlin, Germany, 1996.

[50] C. D. Constantinides, E. Atalar, and E. R. McVeigh, "Signal-to-noise measurements in magnitude images from NMR phased arrays," *Magnetic Resonance in Medicine*, vol. 38, no. 5, pp. 852–857, 1997.

[51] A. Hammers, C.-H. Chen, L. Lemieux et al., "Statistical neuroanatomy of the human inferior frontal gyrus and probabilistic atlas in a standard stereotaxic space," *Human Brain Mapping*, vol. 28, no. 1, pp. 34–48, 2007.

[52] J. V. Allom and P. Coupé, "volBrain: an online MRI brain volumetry system," *Frontiers in Neuroinformatics*, vol. 10, no. 54, pp. 1–14, 2016.

[53] C. R. Jack and Jr.: Alzheimer's disease neuroimaging initiative dataset, http://adni.loni.usc.edu/.

[54] A. Foi, "Noise estimation and removal in MR imaging: the variance-stabilization approach," in *Proceedings of the 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1809–1814, Chicago, IL, USA, March 2011.

[55] A. E. Kavur, N. S. Gezer, M. Barış et al., "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, Article ID 101950, 2021.

[56] A. E. Aslan, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, "CHAOS—Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data," 2019.

[57] J. V. Manjón, P. Coupé, A. Buades et al., "New methods for MRI denoising based on sparseness and self-similarity," *Medical Image Analysis*, vol. 16, no. 1, pp. 18–27, 2012.

[58] S. Louis Collins, K. R. L. Reddy, and D. S. Rao, "Denoising and segmentation of MR images using fourth order non-linear adaptive PDE and new convergent clustering," *International Journal of Imaging Systems and Technology*, vol. 29, no. 3, pp. 195–209, 2019.

[59] O. Oktay, E. Ferrante, K. Kamnitsas et al., "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2018.

[60] Y. Heinrich, B. Zhang, W. Zhao et al., "Magnetic resonance image denoising algorithm based on cartoon, texture, and residual parts," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1405647, 10 pages, 2020.

Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 5520281, 9 pages
https://doi.org/10.1155/2021/5520281

Research Article

SGPNet: A Three-Dimensional Multitask Residual Framework for Segmentation and IDH Genotype Prediction of Gliomas

Yao Wang[1], **Yan Wang**[1,2], **Chunjie Guo**[3], **Shuangquan Zhang**[1], **and Lili Yang**[1,4]

[1]*Key Laboratory of Symbol Computation and Knowledge Engineering, Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[2]*School of Artificial Intelligence, Jilin University, Changchun 130012, China*
[3]*Department of Radiology, The First Hospital of Jilin University, Changchun 130012, China*
[4]*Department of Obstetrics, The First Hospital of Jilin University, Changchun 130012, China*

Correspondence should be addressed to Yan Wang; wy6868@jlu.edu.cn and Lili Yang; ylljlu@jlu.edu.cn

Received 29 January 2021; Revised 25 March 2021; Accepted 2 April 2021; Published 19 April 2021

Academic Editor: Vahid Rakhshan

Glioma is the main type of malignant brain tumor in adults, and the status of isocitrate dehydrogenase (IDH) mutation highly affects the diagnosis, treatment, and prognosis of gliomas. Radiographic medical imaging provides a noninvasive platform for sampling both inter and intralesion heterogeneity of gliomas, and previous research has shown that the IDH genotype can be predicted from the fusion of multimodality radiology images. The features of medical images and IDH genotype are vital for medical treatment; however, it still lacks a multitask framework for the segmentation of the lesion areas of gliomas and the prediction of IDH genotype. In this paper, we propose a novel three-dimensional (3D) multitask deep learning model for segmentation and genotype prediction (SGPNet). The residual units are also introduced into the SGPNet that allows the output blocks to extract hierarchical features for different tasks and facilitate the information propagation. Our model reduces 26.6% classification error rates comparing with previous models on the datasets of Multimodal Brain Tumor Segmentation Challenge (BRATS) 2020 and The Cancer Genome Atlas (TCGA) gliomas' databases. Furthermore, we first practically investigate the influence of lesion areas on the performance of IDH genotype prediction by setting different groups of learning targets. The experimental results indicate that the information of lesion areas is more important for the IDH genotype prediction. Our framework is effective and generalizable, which can serve as a highly automated tool to be applied in clinical decision making.

1. Introduction

Glioma is the main type of malignant brain tumor in adults which accounted for approximately 80% of them, and it can be divided into four grades from I to IV according to the World Health Organization (WHO) [1]. Despite the frequency of gliomas, the histology and molecular etiology are variable even in a single pathology class [2]; hence, recognizing the status is crucial for precision medicine. Isocitrate dehydrogenase (IDH) is a general term for IDH1 and IDH2, and previous studies have proved that the IDH genotype (wild-type or mutation) shows significant impacts on the diagnosis, treatment, and prognosis of glioma patients [3–6]. However, identifying the IDH genotype by a biopsy is an invasive and costly procedure that needs a sample of cells from a patient's lesion, while radiographic medical imaging provides a noninvasive platform for sampling both inter and intralesion heterogeneity of gliomas. Previous research has demonstrated the strong correlation between phenotypes (extracted from medical images) and genotypes (extracted from gene expression files), and the prediction of genotypes from phenotypes becomes a fast-developing research field [7].

At present, there have been constructed high-performance models to predict the genotypes of gliomas patients across medical images. Regarding this task, an effective approach is based on radiomics and machine learning algorithms [8, 9]. Radiomics is a method that extracted lesion-related features from medical images by

experienced radiologists using professional software and data-characterization algorithms [10]. The high-dimensional images' data are well represented by the low-dimensional radiomics features after the processing of radiologists, and using these radiomics features allows researchers to build IDH prediction models more easily. Although the radiomics feature-based models perform well on genotype prediction, they still have some limitations. For example, extracting radiomics features depends on radiologists' judgment is a subjective procedure, and it is also affected by factors of the environment of hardware and software. Different radiologists using different software and algorithms may result in slightly different descriptions of the details of the lesion. Besides, all raw images should be processed before the predicting phase, and the low-dimensional features restrict the models for further investigations. Overall, the model's generalization ability and reproducibility are limited by the high-dependency on manual intervention.

Based on the above observations, researchers introduced deep learning (DL) algorithms into genotype prediction tasks. DL, as a subclass of machine learning (ML), reveals a more powerful learning ability. The annotated data are only required for the training phase, and the well-trained models could receive raw images as input for various tasks. The raw images preserve all the information about the lesions and the organism that allow the models to finish more complex tasks. Chang et al. developed a residual convolutional neural network (CNN) using magnetic resonance (MR) sequence images [11]. However, in Chang's work, the MR sequence images are manually selected from whole 3D brain MR images. To directly handle the 3D brain MR images, Liang et al. developed a 3D DenseNet for IDH genotype prediction of low-grade (grade II and III, known as low-grade gliomas, LGG) and high-grade (grade IV, known as glioblastoma multiform, GBM) gliomas' MR images and achieved an accuracy of 84.6% on the validation dataset [12]. DL algorithms also perform well on automatic segmentation tasks, and previous studies have established many high-performance models to segment lesion areas from medical images [13, 14]. Soltaninejad et al. combined DL and ML algorithms to build superpixel-based and supervoxel-based models for brain tumor segmentation and detection [15]. However, these models are incompetent to predict gene mutation statuses which are also important for the treatment of glioma patients. The attention mechanism is also introduced to improve the performance of segmentation. Although the attention mechanism shows potential to be applied to medical image tasks, it significantly increases the computational complexity of models, especially for the 3D MR images. It means that the attention-based models need more cases, and they are more difficult to be well-trained. Liu et al. developed a multitask model including segmentation of brainstem gliomas and prediction of H3 K27M mutation [16]. The phenotypes of MR images and IDH genotype are both important criteria for gliomas' patients to receive proper medical treatment; however, it still lacks a multitask framework for the segmentation of the lesion areas of gliomas and the prediction of IDH genotype.

The brain MR images contain the details of normal tissues and lesion areas. Both normal tissues and lesion areas may affect the performance of genotype prediction. However, previous research studies only focused on conducting the black-box models for the genotype prediction due to constrained by the single-task model structure, which limits the reliability as a computer-aided tool for diagnosis and treatment. Due to the multitask architecture in the SGPNet, we set up controlled experiments to discuss the influence of lesion areas for IDH genotype prediction by setting different groups of learning targets.

In this paper, we focus on a multitask CNN model to address the challenges of the automatic segmentation of low-grade gliomas (LGG) and glioblastoma multiform (GBM) tumor volumes and the prediction of IDH mutation from MR images (SGPNet). Four types of modalities of MR images including T1, T1Gd, T2, and T2-FLAIR are preprocessed and then fed into the SGPNet, and our model consists of a single backbone with two output blocks, one each for segmentation and IDH status. In order to effectively train such a multitask model, we apply a multiloss function for our network and different learning rates for the different blocks. The experimental results indicate that our model reduces 26.6% classification error rates comparing with previous models on the datasets of Multimodal Brain Tumor Segmentation Challenge (BRATS) and The Cancer Genome Atlas (TCGA) gliomas' databases. In addition, we further study the features of lesion areas which influence the performance of IDH genotype prediction. We believe that these experiments can prove the information of lesions which is important for the IDH genotypes prediction and increase the reliability of the IDH genotypes prediction.

2. Materials and Methods

2.1. Gene Profiles and Medical Images Dataset. In this paper, we used two datasets of The Cancer Genome Atlas (TCGA) and Brain Tumor Segmentation Challenge (BRATS) 2020 databases to conduct our experiments. The genotype-related dataset used in this paper is The Cancer Genome Atlas (TCGA) [17] which provides various gene data types, including gene expression profiling, copy number variation profiling, and so on. More specifically, The TCGA dataset provides four methods to identify gene mutation status in parallel, including MuSE [18], MuTect2 [19], SomaticSniper [20], and VarScan2 [21]. We considered one gene to be in mutation status when more than one of these methods indicated this gene is mutated. The BRATS 2020 dataset [22–24] provides multimodalities brain MR images of LGG and GBM patients, including T1, T1Gd, T2, and T2-FLAIR volumes. One of the sources in the BRATS dataset is The Cancer Imaging Archive (TCIA) dataset [25], which allows us to build cross-referenced MR images and gene expression profiles data according to the project ID in both datasets. The subtypes of the segmentation labels include the necrotic and the nonenhancing (NCR and NET), the peritumoral edema (ED), the enhancing tumor (ET), and the background. In

this paper, considering the scale of the datasets and our research content, we integrate the NCR and NET, ED, and ET into the lesion label, and it can make the evaluation of our experimental results more concise. Totally, 121 cross-referenced patients' data are collected from the above datasets which include 56 mutant cases and 65 wild-type cases, respectively.

2.2. Data Processing. The original MR images have been manually annotated by clinical experts; each entity consists of four modalities volumes (T1, T1Gd, T2, and T2-FLAIR) and the ground-truth segmentation labels, and all those images have the same shape of 155*240*240 pixels. The data preprocessing procedure has the following steps. (1) Every image is cropped to remove the black background. (2) Following the cropped image is reshaped into the unified shape of 144*144*144 pixels, and then all images except for segmentation labels are normalized to zero mean and unit standard deviation. (3) The four modalities are concatenated as four input channels. Figure 1 shows the above steps of the preprocessing procedure. Considering the scale of dataset size, we also apply the data augmentation technique, and the operations of shift and flip are randomly chosen with a fifty percent chance in each training step.

2.3. Model Architecture. In the segmentation task, using low-level details of the input image is proved to be important when the size of datasets is limited; as a result, U-Net has achieved high performance and been widely applied on medical image segmentation [26–28]. Besides, degradation is also a common problem when the network architecture is deep [29]. Inspired by this research, we modify the hyperparameters of 3D U-Net and introduce skip-connection into our model. The basic shape of our framework is based on the standard U-Net containing two paths called contracting path (left side) and expansive path (right side). There are five pairs of blocks employed in the two paths, where the output of the block in the contracting path is concatenated as part of the input of the block in the corresponding expansive path. These connections create a quick pathway for information between high-level and low-level feature maps which is facilitating the gradient backward propagation and compensating finer details into high-level semantic features [30]. Besides, these connections allow the output blocks to extract multilevel features for different tasks from the backbone of SGPNet.

Our proposed network is consisting of one backbone and two output blocks, illustrated in Figure 2. More specifically, [*Conv* (in, out, kernel size, stride)] represents the 3D convolution layer; the items in four-tuple (in, out, kernel size, stride) represent input channels, output channels, kernel size, and stride of the convolutional layer, respectively. IN represents the instance normalization (IN) layer which is designed to remove the instance-specific contrast information from the input image [31], and Up is the up-sampling layer. FC represents the fully connected layer for the prediction of IDH genotype. LR is the following leaky rectified linear unit (LeakyReLU) activation function:

$$\mathrm{LR}: \phi(x) = \begin{cases} x, & \text{if } x > 0, \\ 0.1x, & \text{otherwise.} \end{cases} \tag{1}$$

The segmentation task and the IDH genotype prediction task share most of the weights in the backbone. In general, our network is an end-to-end model, which receives four channels of MR images as input and outputs the segmentation labels and predicted IDH mutation status.

In the contracting path, we replace the max-pooling layer in the standard U-Net, with one 3*3*3 convolution with a stride of 2 for down-sampling and double the number of output channels, followed by two repeated 3*3*3 convolutions with a stride of 1. LR and IN are also added after the convolution layer. The blue dotted line represents the skip-connection; it adds the output of the first convolution layer with the output of the last convolution layer in each block. In the expansive path, the input of each block is the concatenation of the previous block and the corresponding feature map from the paired contracting path. The first 3*3*3 convolution integrates the information of concatenated input, followed by a 1*1*1 convolution that halves the number of input channels. The upsampling layer follows these two convolutions and uses the nearest neighbor interpolation algorithm to double the width and height of the input features, followed by a 3*3*3 convolution to further half the number of input channels. These two output blocks have also introduced the idea of skip-connection. For the segmentation and IDH genotype output blocks, the input of these blocks is from three different levels' blocks in the expansive path of the backbone.

2.4. Evaluation Metrics. In this section, we use four metrics to assess SGPNet including specificity (SP), sensitivity (SN), accuracy (ACC), and area under the receiver operating characteristic curve (AUC) for IDH status prediction task and dice similarity coefficient (DSC) for segmentation task. Specificity (SP) measures the proportion of negatives that are correctly predicted, as in equation (2), and sensitivity (SN) is the measurements of true positive rate, as in equation (3). ACC is the fraction of the total samples that are identified correctly, as in equation (4). AUC calculates the probability that a randomly selected positive example ranked above a randomly selected negative one. Dice similarity coefficient (DSC) is designed to score how closely the predicted segmentation labels matched the annotated ground-truth segmentation labels, as in equation (5).

$$SP = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

$$SN = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

$$ACC = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}, \tag{4}$$

$$DSC = \frac{2 * \mathrm{TP}}{(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TP} + \mathrm{FN})}. \tag{5}$$

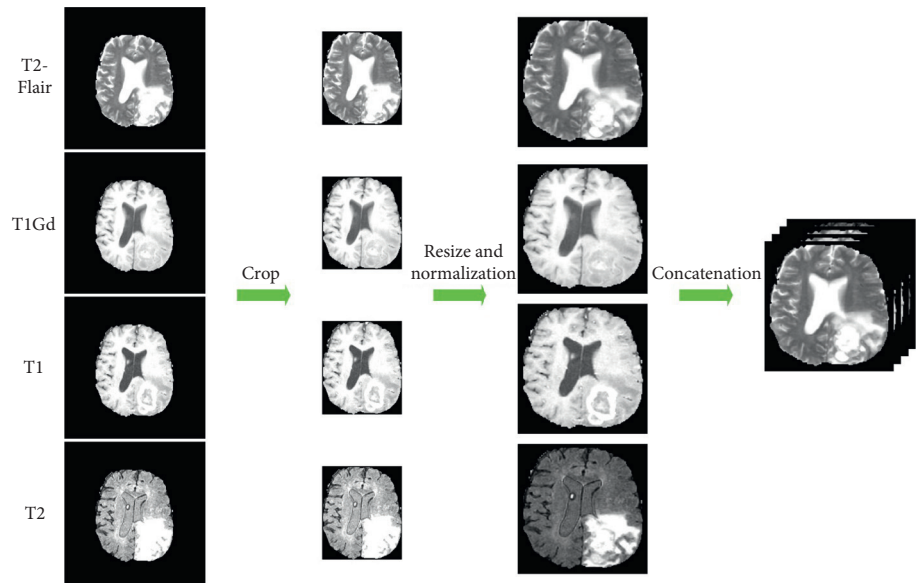


Figure 1: The diagram of our image preprocessing procedure of four modalities volumes (T2-Flair, T1Gd, T1, and T2) which includes following steps. Crop: remove the black background; Resize and Normalization: each modality volume is resized into a unified shape and then normalized to zero mean and unit standard deviation; Concatenation: each modality volume is concatenated into one image.

There are four definitions introduced to calculate the above items: true positive (TP) is the quantity of the correctly predicted positive class, likewise, true negative (TN) is the number of correctly predicted negative class. False positive (FP) is the quantity of incorrectly predicted positive class, and false negative (FN) is the quantity of incorrectly predicted negative class.

2.5. Implementation Details. Considering the evaluation metrics, cross-entropy and dice loss are the objective functions of our network. In the task of gene mutation prediction, the IDH status is encoded into two labels (wild-type and mutation). The binary cross-entropy (BCE) loss function $\mathscr{L}_1$ is used to calculate the similarity between the predicted labels and ground-truth labels, which is defined as follows:

$$\mathscr{L}_1 = -\sum y \log\widehat{y} + (1-y)\log(1-\widehat{y}), \tag{6}$$

where $\widehat{y}$ represents the model's prediction of class possibilities and y represents the ground-truth labels.

The dice loss function is aimed to calculate the spatial overlap accuracy of predicted segmentation labels compared with manually annotated labels which are defined as follows:

$$\mathscr{L}_2 = 1 - \mathrm{DSC} = 1 - \frac{2 * \mathrm{TP}}{(\mathrm{TP}+\mathrm{FP}) + (\mathrm{TP}+\mathrm{FN})}. \tag{7}$$

The ground-truth segmentation labels contain more information than the IDH mutation status, so it may be not ideal to weigh segmentation error equally with classification error. In order to integrate the above loss functions, we define the total loss as follows:

$$L = \mathscr{L}_1 + k * \mathscr{L}_2, \tag{8}$$

where k is the parameter to balance the segmentation error and classification error. In order to dynamically balance the dice loss and classification loss, the parameter k in the total loss function is defined as $(\mathscr{L}_2/\mathscr{L}_1 + \mathscr{L}_2)$, so the total loss function can be given by the following formula:

$$L = \mathscr{L}_1 + \frac{\mathscr{L}_2}{\mathscr{L}_1 + \mathscr{L}_2} * \mathscr{L}_2. \tag{9}$$

We set different learning rates for different parts of our network. In particular, the learning rate is set to 0.0001 for the backbone and segmentation labels output block, and it is set to 0.00005 for the IDH status prediction block. Moreover, we adopt learning rate scheduling with cosine annealing during the training phase. The weights of our network are optimized by the Adam [32] method with a minibatch size of two.

3. Experiments and Results

In this section, we present a series of experiments to demonstrate the performance of the proposed multitask model; we test SGPNet on the BRAST and TCGA datasets and compare SGPNet with three existing models. Furthermore, we discuss the impact of the lesion's information for the IDH status prediction task. Overall, 121 gliomas cases are involved including 56 mutant cases and 65 wild-type cases. The reproducibility of the results is verified in fivefold cross-validations, and the final results are the average of the cross-validations.

3.1. Multitask Model for Segmentation and IDH Genotype Prediction. In order to evaluate the performance of our proposed model, we compare SGPNet with three different models. ACC, SE, SP, and AUC metrics are utilized to

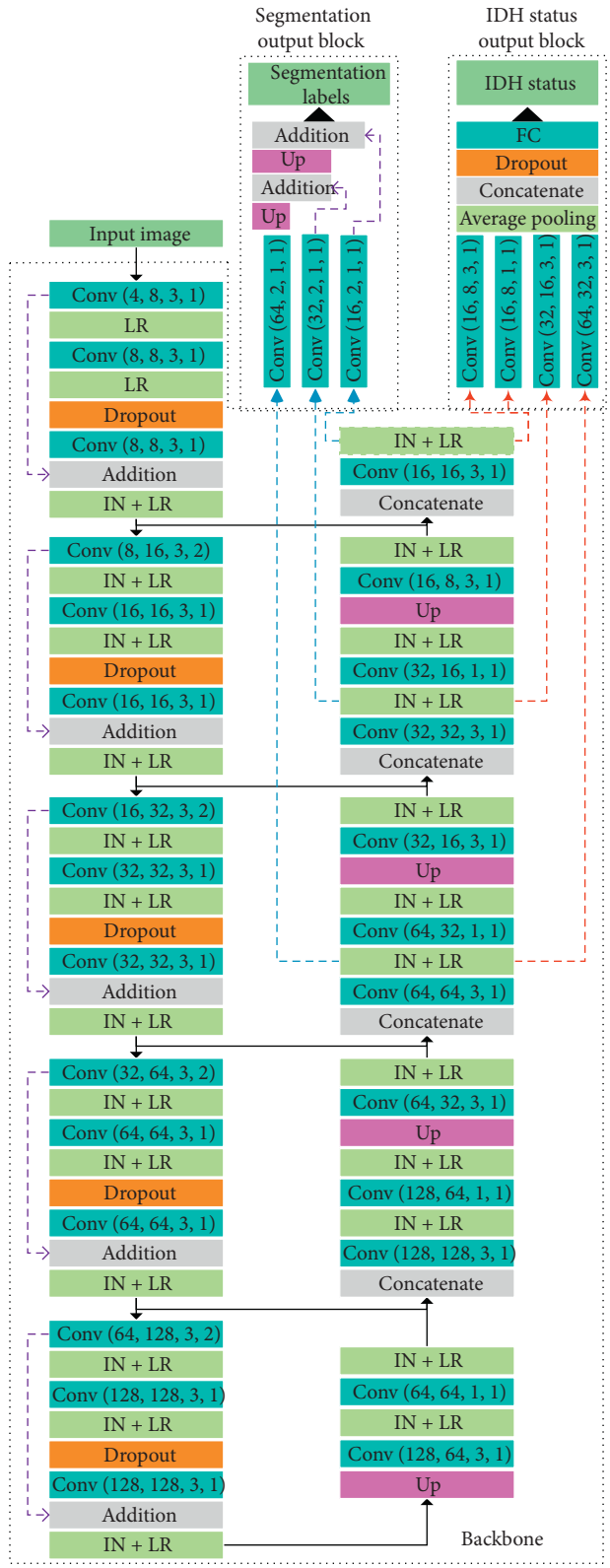


Figure 2: Architecture of the SGPNet and the details of parameters. The SGPNet consists of a single backbone with two output blocks, one each for segmentation and IDH status. The backbone contains two paths called contracting path (left side) and expansive path (right side).

quantitatively evaluate the performance of the prediction of IDH genotype, and the DSC metric is used to evaluate the performance of the segmentation task. Table 1 shows the ACC, SN, SP, AUC, and DSC of all models on the performance of the IDH genotype prediction task and segmentation task. Figure 3 illustrates the qualitative

Table 1: Comparisons of the proposed and other deep learning-based models on cross-referenced TCGA and BRATS datasets.

Method	ACC	SN	SP	AUC	DSC
Liang et al. [12]	0.846	0.785	0.880	0.857	—
Chang et al. [11]	0.857	—	—	0.940	—
3D U-Net [27]	—	—	—	—	0.920
SGPNet (only segment)	—	—	—	—	0.937
SGPNet (multitasks)	0.895	0.907	0.883	0.949	0.935

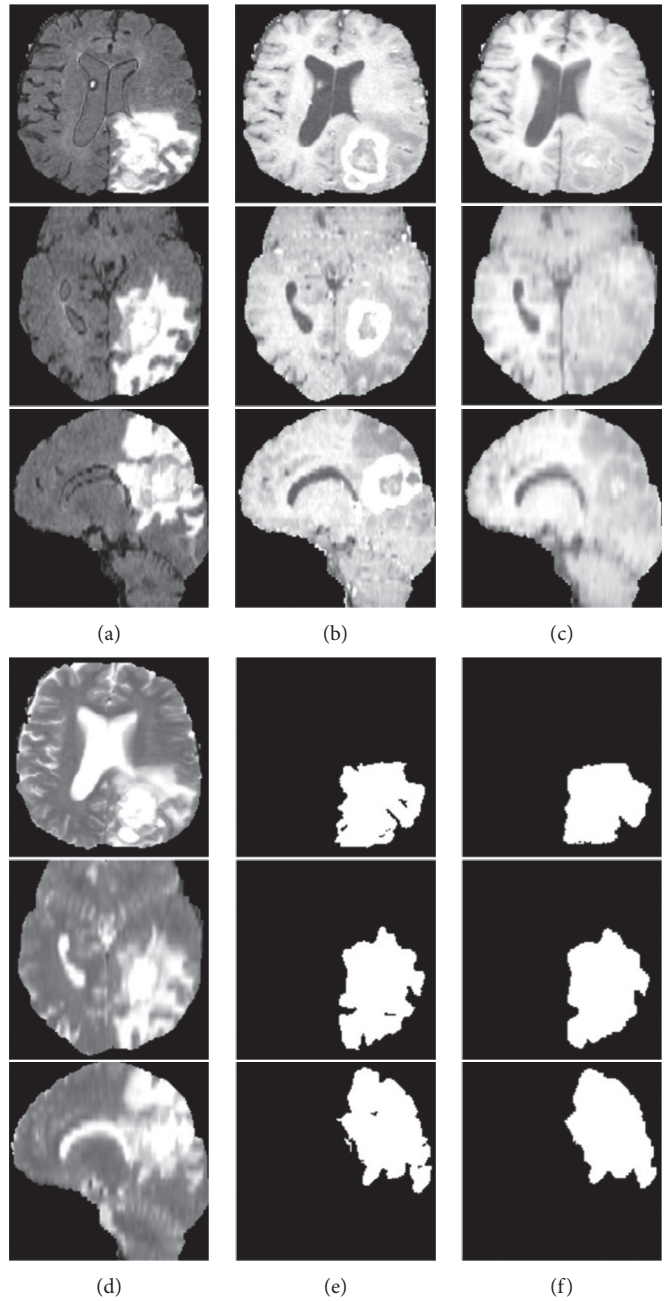


Figure 3: Illustration of the performance of segmentation task of the SGPNet: (a–d) the 3D T2-Flair, T1Gd, T1, and T2 volumes of a glioma patient in the axial, coronal, and sagittal slices after data preprocessing; (e) the ground-truth segmentation labels of the lesion; (f) the segmentation results are predicted by the SGPNet.

segmentation results of lesion areas with our SGPNet, which demonstrates that the SGPNet can determine the boundary of the lesion accurately.

Different from single-task segmentation and classification models, the SGPNet not only can segment the lesions of gliomas but also predicts the IDH genotype depending on the brain MR images. The positive predictive value (PPV) and negative predictive value (NPV) of the SGPNet achieve 0.894 and 0.908, respectively. Moreover, these experimental results show that our proposed model reduced 26.6% classification error rates compared with previous models and performed well on gliomas' lesions segmentation.

3.2. The Comparisons with Different Groups of Learning Targets. The brain MR images contain the details of normal tissues and lesion areas. Both normal and lesion areas may possibly influence genotype prediction. The multitask model structure allows us to set different groups of learning targets to investigate if the information of lesion areas or the whole-brain MR images may be more likely to influence the genotype prediction, which might increase the reliability as a computer-aided tool for diagnosis and treatment. In this section, we carry out three controlled experiments for analysing the relationship between the genotypes and phenotypes by training SGPNet with different groups of learning targets: (1) SGPNet is only trained with IDH genotype; (2) SGPNet is trained with ground-truth segmentation labels and IDH genotype; and (3) SGPNet is trained with randomly generated tensor as segmentation labels and IDH genotype. Table 2 shows the performance of IDH genotype prediction across three controlled experiments. Figure 4 compares the comparative ROC curves of different experiments.

The total loss function is simplified as a single-task objective function $\mathscr{L}_1$ when SGPNet is only trained with IDH genotype labels. After that, SGPNet is considered as a classifier of IDH genotype, and the performance of SGPNet is worse than Liang et al. and Chang et al. [11, 12]. One important reason is that Liang et al. and Chang et al. crop the lesion areas as the models' input, while our model receives whole-brain MR images as input, which increases the difficulty for the model to extract useful features considering the limited information of IDH genotype labels. When the ground-truth segmentation labels are added as learning targets, the performance of the model is significantly improved. However, the first experiment uses a single-task objective function $\mathscr{L}_1$, while the second experiment uses the multitask objective function $\mathscr{L}$. To further discuss the influence of the objective function, we set up the third experiment that regards randomly generated segmentation labels as learning targets. It means that the segmentation output block learns the wrong features of lesion areas while the IDH status output block can still learn the features of the whole MR images; as a result, the performance of the model is significantly cut down. After comparing these experimental results, we can infer that the ground-truth segmentation labels promote the performance of IDH genotype prediction, and the lesions information is more important to predict the IDH genotype.

4. Discussion

Developing an automatic segmentation of 3D gliomas lesion is a challenging task, considering the wide variability in tumor size, form, and strength. Furthermore, the mutation status of IDH can be used as a qualified biomarker for selecting diagnostic and therapeutic approaches for gliomas patients. Previous studies have focused on the prediction of genotypes from medical images [8, 9, 11, 12]; however, these single-task models show the limitation of their practicality and scalability. However, it still lacks a multitask model for segmentation and IDH genotype prediction of gliomas. Besides, there is no research to compare the influence of the images' features of whole MR images and lesion areas to the prediction of IDH genotype.

The SGPNet is an end-to-end framework designed to address the challenges of segmentation and IDH genotype prediction of gliomas. In Section 3.1, the experimental results indicate the significant improvement of the performance of IDH genotype prediction, and the prediction error rates reduce 26.6%, comparing to the models of Liang et al. and Chang et al. [11, 12]. Due to the multitask model architecture, in Section 3.2, we further discuss if the information of gliomas' lesions or whole MR images is more likely to affect the prediction of IDH genotype by setting different learning target groups. The experimental results indicate that providing the ground-truth segmentation labels as learning targets will promote the performance of IDH genotype prediction comparing with other experiments. Overall, we infer that the information of lesion areas is more important for IDH genotype prediction, which increases the reliability as a computer-aided tool for diagnosis and treatment.

In clinical practice, the diagnosis of glioma is usually made by experts based on the various MR images and gene mutation statuses. The different modalities of MR images can reflect different characteristics of the lesions. For example, T1 provides anatomical information, and T2 is sensitive to the edema area and reflects the morphological information of tumors [33]. The SGPNet can integrate multimodality MR images to predict the boundary of lesion areas and the IDH genotype of the patients, and it can reduce doctors' workload and help doctors to choose the proper treatment for the patients. The SGPNet is feasible for segmentation and genotype prediction because the backbone of our framework is designed to learn the intrinsic information of patients' lesions. Meanwhile, the framework of SGPNet can be used to segment other tissue lesions or predict other genotypes when it is well-trained on the corresponding dataset. The SGPNet can be also applied to multicenters and larger-scale multi-sequence MR image datasets because the backbone in our models is generic for any MR image collected from different institutions, equipment, and modalities. Moreover, increasing the scale of training datasets can improve the generalization ability of the SGPNet. Generating probability density distributions for different tissue types is also an effective approach to reduce noise reduce environmental noise and improve generalization ability [34]. Therefore, the design of an automatic multitask model for

Table 2: ACC, SN, and SP of SGPNet across different groups of learning targets.

Learning targets	ACC	SN	SP	AUC
Only IDH genotype	0.803	0.799	0.807	0.825
IDH genotype + segmentation labels (baseline)	0.895	0.907	0.883	0.949
IDH genotype + random generated segmentation labels	0.720	0.642	0.780	0.727

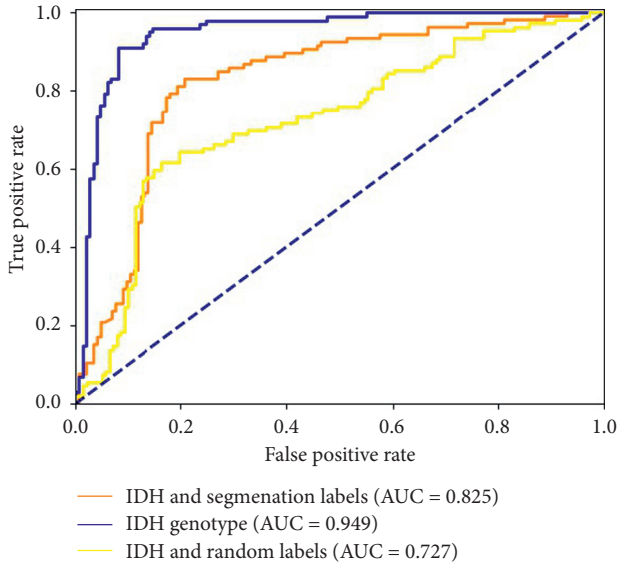


Figure 4: The ROC curves of the SGPNet across different groups of learning targets, which indicate that the SGPNet performs best when the ground-truth segmentation labels are provided as the learning target.

gliomas has superior clinical value. In the future, we will further develop our framework and apply the SGPNet to more types of diseases and genes.

5. Conclusion

In this paper, we present a novel multitask 3D framework named SGPNet for automatic segmentation of gliomas lesions and prediction of IDH mutation status from MR images. Our framework employs a backbone for learning the intrinsic MR image information, two output blocks for segmentation and IDH genotype prediction of gliomas. The experimental results indicate that our architecture achieves a better IDH genotype prediction performance on public TCGA and BRATS 2020 datasets comparing with previous studies and achieves a good result on the segmentation task. Furthermore, we compare the influence of the images' features of whole MR images and lesion areas to the prediction of genotype and the experimental results, indicating that the information of patients' lesions is more significant for the prediction of IDH genotype. In summary, the accurate segmentation of glioma lesion regions and prediction of IDH mutation status will improve therapeutic criteria and assist doctors in diagnosis and treatment.

Data Availability

The MRI data used to support the findings of this study have been deposited in the BRATS repository (http://braintumorsegmentation.org/), and the gene profiles data used to support the findings of this study have been deposited in the TCGA-GBM and TCGA-LGG repositories (https://portal.gdc.cancer.gov/projects/TCGA-GBM and https://portal.gdc.cancer.gov/projects/TCGA-LGG).

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 62072212 and 81600923), the Development Project of Jilin Province of China (nos. 20200401083GX and 2020C003), and the Foundation of Health and Family Planning Commission of Jilin Province (no. 2020J052). This work was also supported by the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (no. 20180622002JC).

References

[1] D. N. Louis, A. Perry, G. Reifenberger et al., "The 2016 world health organization classification of tumors of the central nervous system: a summary," *Acta Neuropathologica*, vol. 131, no. 6, pp. 803–820, 2016.

[2] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer Genetics*, vol. 205, no. 12, pp. 613–621, 2012.

[3] C. Hartmann, B. Hentschel, W. Wick et al., "Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than IDH1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas," *Acta Neuropathologica*, vol. 120, no. 6, pp. 707–718, 2010.

[4] S. Nobusawa, T. Watanabe, P. Kleihues, and H. Ohgaki, "IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas," *Clinical Cancer Research*, vol. 15, no. 19, pp. 6002–6007, 2009.

[5] A. Olar, K. M. Wani, K. D. Alfaro-Munoz et al., "IDH mutation status and role of WHO grade and mitotic index in overall survival in grade II-III diffuse gliomas," *Acta Neuropathologica*, vol. 129, no. 4, pp. 585–596, 2015.

[6] H. Yan, D. W. Parsons, G. Jin et al., "IDH1 and IDH2 mutations in gliomas," *New England Journal of Medicine*, vol. 360, no. 8, pp. 765–773, 2009.

[7] Y. Wang, Y. Wang, C. Guo et al., "Cancer genotypes prediction and associations analysis from imaging phenotypes: a survey on radiogenomics," *Biomarkers in Medicine*, vol. 14, no. 12, pp. 1151–1164, 2020.

[8] H. Arita, M. Kinoshita, A. Kawaguchi et al., "Lesion location implemented magnetic resonance imaging radiomics for predicting IDH and TERT promoter mutations in grade II/III gliomas," *Scientific Reports*, vol. 8, no. 1, Article ID 11773, 2018.

[9] B. Zhang, K. Chang, S. Ramkissoon et al., "Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas," *Neuro-Oncology*, vol. 19, no. 1, pp. 109–117, 2017.

[10] P. Lambin, E. Rios-Velazquez, R Leijenaar et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer (Oxford, England: 1990)*, vol. 48, no. 4, pp. 441–446, 2012.

[11] K. Chang, H. X. Bai, H. Zhou et al., "Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging," *Clinical Cancer Research*, vol. 24, no. 5, pp. 1073–1081, 2018.

[12] S. Liang, R. Zhang, D. Liang et al., "Multimodal 3D DenseNet for IDH genotype prediction in gliomas," *Genes*, vol. 9, no. 8, p. 382, 2018.

[13] T. Hassanzadeh, D. Essam, and R. Sarker, "2D to 3D evolutionary deep convolutional neural networks for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 712–721, 2021.

[14] Z. Gao, H. Zhang, S. Dong et al., "Salient object detection in the distributed cloud-edge intelligent network," *IEEE Network*, vol. 34, no. 2, pp. 216–224, 2020.

[15] M. Soltaninejad, L. Zhang, T. Lambrou et al., "MRI brain tumor segmentation and patient survival prediction using random forests and fully convolutional networks," in *Proceedings of the International MICCAI Brainlesion Workshop*, pp. 204–215, Quebec City, Canada, September 2017.

[16] J. Liu, F. Chen, C. Pan et al., "A cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1943–1952, 2018.

[17] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): an immeasurable source of knowledge," *Contemporary Oncology*, vol. 19, no. 1A, pp. A68–A77, 2015.

[18] Y. Fan, L. Xi, D. Hughes et al., "MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data," *Genome Biology*, vol. 17, no. 1, p. 178, 2016.

[19] A. McKenna, M. Hanna, E. Banks et al., "The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[20] D. E. Larson, C. C. Harris, K. Chen et al., "SomaticSniper: identification of somatic point mutations in whole genome sequencing data," *Bioinformatics*, vol. 28, no. 3, pp. 311–317, 2012.

[21] D. C. Koboldt, Q. Zhang, D. E. Larson et al., "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Research*, vol. 22, no. 3, pp. 568–576, 2012.

[22] S. Bakas, H. Akbari, A. Sotiras et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, no. 1, Article ID 170117, 2017.

[23] B. H. Menze, A. Jakab, S. Bauer et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[24] S. Bakas, M. Reyes, A. Jakab et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, https://arxiv.org/abs/1811.02629.

[25] K. Clark, B. Vendt, K. Smith et al., "The cancer imaging archive (TCIA): maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Munich, Germany, October 2015.

[27] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: learning dense volumetric segmentation from sparse annotation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, Athens, Greece, October 2016.

[28] T. Zhou, S. Ruan, and S. Canu, "A review: deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3-4, Article ID 100004, 2019.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[30] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: the missing ingredient for fast stylization," 2016, https://arxiv.org/abs/1607.08022.

[32] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," 2015, https://arxiv.org/abs/1412.6980.

[33] Y. Wu, Z. Zhao, W. Wu, Y. Lin, and M. Wang, "Automatic glioma segmentation based on adaptive superpixel," *BMC Medical Imaging*, vol. 19, no. 1, p. 73, 2019.

[34] F. Raschke, T. R. Barrick, T. L. Jones et al., "Tissue-type mapping of gliomas," *NeuroImage: Clinical*, vol. 21, Article ID 101648, 2019.

Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 5580914, 17 pages
https://doi.org/10.1155/2021/5580914

Research Article

Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images Using Transfer Learning with EfficientNets

Chanaleä Munien and Serestina Viriri

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 217013433, South Africa

Correspondence should be addressed to Serestina Viriri; viriris@ukzn.ac.za

Received 23 January 2021; Revised 15 March 2021; Accepted 29 March 2021; Published 9 April 2021

Academic Editor: Vahid Rakhshan

Breast cancer is a fatal disease and is a leading cause of death in women worldwide. The process of diagnosis based on biopsy tissue is nontrivial, time-consuming, and prone to human error, and there may be conflict about the final diagnosis due to interobserver variability. Computer-aided diagnosis systems have been designed and implemented to combat these issues. These systems contribute significantly to increasing the efficiency and accuracy and reducing the cost of diagnosis. Moreover, these systems must perform better so that their determined diagnosis can be more reliable. This research investigates the application of the EfficientNet architecture for the classification of hematoxylin and eosin-stained breast cancer histology images provided by the ICIAR2018 dataset. Specifically, seven EfficientNets were fine-tuned and evaluated on their ability to classify images into four classes: *normal, benign, in situ carcinoma,* and *invasive carcinoma*. Moreover, two standard stain normalization techniques, Reinhard and Macenko, were observed to measure the impact of stain normalization on performance. The outcome of this approach reveals that the EfficientNet-B2 model yielded an accuracy and sensitivity of 98.33% using Reinhard stain normalization method on the training images and an accuracy and sensitivity of 96.67% using the Macenko stain normalization method. These satisfactory results indicate that transferring generic features from natural images to medical images through fine-tuning on EfficientNets can achieve satisfactory results.

1. Introduction and Background

One of the leading causes of death in women throughout the world is breast cancer [1]. It is defined as a group of diseases in which cells within the tissue of the breast alter and divide in an uncontrolled manner, generally resulting in lumps or growths. This type of cancer often begins in the milk glands or ducts connecting these glands to the nipple. In the beginning stages of the illness, the small tumour that appears is much easier to treat effectively, averting the disease's progression and decreasing the morbidity rates; this is why screening is crucial for early detection [2].

The process of breast cancer diagnosis begins with palpation, periodic mammography, and ultrasonic imaging inspection. The results of these procedures indicate whether further testing is required. If cancer is suspected in a patient, a biopsy is performed and tissue for microscopic analysis is procured so that a pathologist may conduct a histological examination of the extracted tissue to confirm the diagnosis [2, 3]. Once the biopsy is complete, the tissue is analyzed in a laboratory. The tissue preparation process must begin with formalin fixation and, after that, embedding in paraffin sections. The paraffin blocks are then sliced and fixed on glass slides. Unfortunately, interesting structures such as the cytoplasm and nuclei in the tissue are not yet apparent at this point. The lack of clarity in the tissue necessitates staining of the tissue so that the structures can become more visible. Typically, a standard and well-known staining protocol, using hematoxylin and eosin, is applied. When added to the tissue, the hematoxylin can bind itself to deoxyribonucleic acid, which results in the nuclei in the tissue being dyed a blue/purple color. On the other hand, the eosin can bind

itself to proteins, and, as a result, other relevant structures such as the stroma and cytoplasm are dyed a pink color. Traditionally, after staining, the glass slide is coverslipped and forwarded to a pathologist for examination [4]. Routinely, the expert gathers information on the texture, size, shape, organization, interactions, and spatial arrangements of the nuclei. Additionally, the variability within, density of, and overall structure of the tissue is analyzed. In particular, the information concerning the nuclei features is relevant for distinguishing between noncarcinoma and carcinoma cells. In contrast, the information concerning the tissue structure is relevant for distinguishing between *in situ* and invasive carcinoma cells [5].

The noncarcinoma class consists of normal tissue and benign lesions; these tissues are nonmalignant and do not require immediate medical attention. *In situ* and invasive carcinoma, on the other hand, are malignant and become continuously more lethal without treatment. Specifically, *in situ* carcinoma refers to the presence of atypical cells that are confined to the layer of tissue in the breast from which it stemmed. Invasive carcinoma refers to the presence of atypical cells that invades the surrounding normal tissue, beyond the glands or ducts from where the cells originated [2]. Invasive carcinoma is complicated to treat, as it poses a risk to the entire body [3]. This threat means that the odds of surviving this level of cancer decreases as the progression stages increase. Moreover, without proper and adequate treatment, a patient's *in situ* carcinoma tissue can develop into invasive carcinoma tissue. Therefore, it is of paramount importance that biopsy tissue is examined correctly and efficiently so that a diagnosis can be confirmed and, subsequently, treatment can begin. Examples of histology images belonging to each of these classes are shown in Figure 1.

The task of performing a practical examination on the tissue is not simple and straightforward. On the contrary, it is rather time-consuming and, above all, prone to human error. The average diagnostic accuracy between professionals is around 75% [6]. These issues can result in severe and fatal consequences for patients who are incorrectly diagnosed [7].

The advancement of image acquisition devices that create whole slide images (WSI) from scanning conventional glass slides has promoted digital pathology [8]. The field of digital pathology focuses on bringing improvement in accuracy and efficiency to the pathology practice [9] by associating histopathological analysis with the study of WSI [8].

An excellent solution to address the limitations of human diagnosis is computer-aided diagnosis (CAD) systems, which are developed to automatically analyze the WSI and provide a potential diagnosis based on the image. These systems currently contribute to improving efficiency and reducing both the cost of diagnosis and interobserver variability [5, 10]. Even though current CAD systems that operate at high sensitivity provide relatively good performance, they will remain a second-opinion clinical procedure until the performance is significantly improved [10].

Recently, deep learning approaches to the development of CAD systems have produced promising results. Previous attempts to classify breast cancer histology images using a combination of handcrafted feature extraction methods and traditional machine learning algorithms required additional knowledge and were time-consuming to develop. Conversely, deep learning methods automate this process. These systems allow pathologists to focus on difficult diagnosis cases [11].

Hence, to ensure early diagnosis in breast cancer candidates, increase treatment success, and lower mortality rates, early detection is imperative. Although the advent of and advancements in computer-aided systems have benefited the medical field, there is plenty of room for improvement.

1.1. Research Problem. In general, the shortage of available medical experts [12], the time-consuming quest to reach a final decision on a diagnosis, and the issue of interobserver variability justify the need for a system that can automatically and accurately classify breast cancer histopathology images. Previous approaches to this problem have been relatively successful considering the available data and return adequate classification accuracies but tend to be computationally expensive. Thus, this work will explore the use of seven lightweight architectures within the EfficientNet family [13]. Since the EfficientNet models were designed to optimize available resources, while maintaining high accuracies, a CAD system that performs at the level of the current state-of-the-art deep learning approaches, while consuming less space and training time, is desirable. Transfer learning techniques have become a popular addition to deep learning solutions for classification tasks. In particular, many state-of-the-art approaches utilize fine-tuning to enhance performance [14]. Therefore, this research explores the application of seven pretrained EfficientNets for the classification of breast cancer histology images. Furthermore, the addition of stain normalization to the preprocessing step will be evaluated. Hence, the primary question that this research will answer is, "Can fine-tuned EfficientNets achieve similar results to current state-of-the-art approaches for the application of classifying breast cancer histology images?"

1.2. Research Contributions. In this research, the application of seven versions of EfficientNets with transfer learning for breast cancer histology image classification is investigated. The proposed architecture was able to effectively extract and learn the global features in an image, such as the tissue and nuclei organization. Of the seven models tested, the EfficientNet-B2 architecture produced superior results with an accuracy of 98.33% and sensitivity of 98.44%.

The key takeaway from this investigation is that the simple and straightforward approach to using EfficientNets for the classification of breast cancer histology images reduces training time while maintaining similar accuracies to previously proposed computationally expensive approaches.

1.3. Paper Structure. The remainder of the paper is structured as follows: Section 2, the literature review, provides details on previous successful approaches. Section 3, the

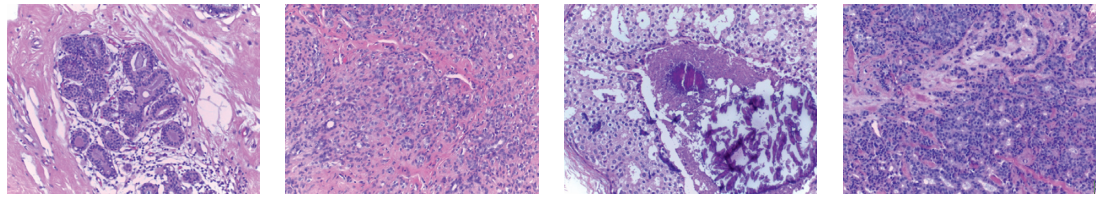


Figure 1: Different classes of histology microscopy images: (a) normal; (b) benign; (c) in situ; (d) invasive.

methods and techniques, provides insight into the framework followed in this study. Section 4, the results, provides details of the results that were obtained during the research. Finally, section 5 elaborates on the insights of this work and concludes the paper.

2. Literature Review

Currently, computer-aided diagnosis (CAD) systems occupy the position of aiding physicians during the process of diagnosis, by easing their workload and reducing the disagreement that stems from the subjective interpretation of pathologists. However, the performance of these systems must be enhanced before they can be considered more dependable than a second-opinion system [10].

2.1. Traditional Approaches. In the traditional approach, expert domain knowledge is required so that the correct features may be handcrafted; this is a time-consuming endeavor. Nevertheless, the approach yields acceptable results on the datasets used. For instance, Kowal [15] used multiple clustering algorithms to achieve nuclei segmentation on microscopic images. Segmentation made it possible to extract microscopic, textural, and topological features so that classifiers could be trained and images could be classified as either benign or malignant. The accuracy of patient-wise classification was in the range of 96–100%. It is worth noting that this method performs poorly when an image contains overlapping nuclei or a small number of nuclei. In this case, either the approach fails to identify the nuclei or the clustering algorithms could return unreliable results. Therefore, in order to attain an acceptable detection accuracy, a large number of sample images are required. Hence, it is evident that accurate nuclei segmentation is not a straightforward task; this can also be attributed to the variability in tissue appearance or the presence of clustered or tightly clumped nuclei [5].

An alternative approach is utilizing information on tissue organization as in the work by Belsare et al. [16], which presents a framework to classify images into malignant and nonmalignant. Firstly, segmentation was done using spatio-color-texture graphs. After that, statistical feature analysis was employed, and classification was achieved with a linear discriminant classifier. The choice of this classifier considerably impacted the outcome of this approach, as the result outperformed the use of k-nearest-neighbor and state vector machine classifiers, especially for the detection of nonmalignant tissue. Accuracies of 100% and 80% were achieved for nonmalignant and malignant images, respectively.

2.2. Deep Learning Approaches. The increase in the availability of computing power has led to the emergence of advanced architectures called convolutional neural networks (CNNs). Contrary to the conventional approach, no expert domain knowledge is required to define algorithms for segmentation, feature extraction, and classification, but instead expert knowledge is needed to annotate the dataset for a CNN to achieve superior results. Instead, these networks can automatically determine and extract discriminative features in an image that contribute to the classification of the image. Generally, a CNN will use a training set of images to learn features that are unique to each class so that when a similar feature is detected in an unseen image, the network will be able to assign the image to a class with confidence.

2.2.1. Convolutional Neural Network Approaches. The success of convolutional neural networks (CNNs) with general computer vision tasks motivated researchers to employ these models for classifying histopathology images. For the classification of hematoxylin and eosin-stained breast cancer histology images, both Araújo et al. [5] and Vo et al. [17] used the Bioimaging 2015 dataset [18] and classified the images into four classes (normal, benign, *in situ*, and invasive) and two groups (carcinoma and noncarcinoma). The former work [5] proposes a CNN that can integrate information from multiple histological scales. The process begins with stain normalization via the method proposed by Macenko et al. [19] in a bid to correct color discrepancies. After that, 12 512×512 overlapping patches were extracted from each image. The chosen size of the patches ensures that no relevant information is lost during extraction and, therefore, every patch can be appropriately labeled. Then, data augmentation was used to increase the number of images in the dataset. Finally, a patch-wise trained CNN and a fusion of a CNN and support vector machine classifier (CNN + SVM) were used to determine the patch class probability. Image-wise classification was attained through a patch probability fusion method. The evaluation showed that using majority voting strategy as the fusion method produced the best results. Considering all four classes, patch-wise classification with the CNN achieved an accuracy rate of

66.7%, while the CNN + SVM achieved an accuracy rate of 65%. The image-wise classification achieved higher results at 77.8% accuracy for both classifiers. With only two classes, patch-wise accuracy for the CNN was 77.6%, and for the CNN + SVM, the approach yielded 76.9% accuracy. The image-wise classification for the 2-class task produced the best results at 80.6% for the CNN and 83.3% for the CNN + SVM. The reason for the lower patch-wise classification is that images may contain sections of normal-looking tissue. Since during patch generation the extracted patches inherit the image's label, this may confuse the CNN. The increase in image-wise classification accuracy is due to the fusion method that is applied. The authors also recorded the sensitivity rates for each of the classes. It is worth noting that overall, for image-wise classification, the approach was more sensitive to the carcinoma class than the noncarcinoma class. This outcome, although not ideal, is preferable since the architecture that was proposed focuses on correctly classifying the carcinoma (malignant) instances [5].

The approach taken by Vo and Nguyen [17] proposed a combination of an ensemble of deep CNNs and gradient boosting tree classifiers (GBTCs). Stain normalization via Macenko et al. [19] and data augmentation were the initial steps of the process. Unlike the standard data augmentation method of rotating and flipping images, the proposed method [17] incorporates reflection, translation, and random cropping of the images. The normalized and augmented data was then used to train the proposed architecture. Specifically, three deep CNNs (Inception-ResNet-v2) were trained using three different input sizes: 600×600, 450×450, and 300×300. Then, visual features were extracted and fed into GBTCs, which increased classification performance. The majority voting strategy was used to merge the outputs of the GBTCs, resulting in a much more robust solution. Recognition rates of 96.4% for the 4-class classification and 99.5% for the 2-class classification were reported. This result surpasses state-of-the-art achievements. An interesting note is that the authors added global average pooling layers in place of dense (fully connected) layers, and this did not negatively impact the accuracy of the ensemble. Similar to Araújo et al. [5], the authors of this work recorded the sensitivities of their approach. The results indicate that, for the 4-class task, the proposed method struggles with the classification of the *in situ* instances, while the other three classes have incredibly high sensitivities. For the 2-class task, the approach yields a 100% sensitivity on carcinoma instances and 98.9% on noncarcinoma instances. These results indicate that the approach was able to successfully learn both local and global features for the multiclass and binary classification. However, the downfall of this approach is the computational expense.

2.2.2. Convolutional Neural Network with Transfer Learning Approaches. For the TK-AlexNet proposed by Nawaz et al. [3] to classify breast cancer histology images, the classification layers of the AlexNet architecture were replaced with a single convolutional layer, and a max-pooling layer was added before the three fully connected layers with 256, 100, and 4 neurons, respectively. The input size of the proposed network [3] was increased to 512×512. The transfer learning technique used in this application was to fine-tune the last three layers on the ICIAR2018 dataset after having the entire network trained on the ImageNet dataset. The images were stain-normalized with the method proposed in Macenko et al. [19]. An interesting fact is that the authors compared the performance of the model with both non-stain-normalized and stain-normalized images and concluded that using the latter resulted in a gain in performance. After that, data augmentation techniques such as mirroring and rotation were applied, and overlapping patches of size 512×512 were extracted from each image. Hence, there was a total of 38400 images generated. Evaluation of the model was done using a train-test split of 80%–20%.

The image-wise accuracy reported in [3] was 81.25%, and the patch-wise accuracy was 75.73%. A noteworthy observation is that the normal and benign classes were classified with 85% sensitivity; however, the *in situ* and invasive carcinoma classes were classified with 75% sensitivity. For a model to be practical as a second-opinion system, it should ideally have a higher sensitivity to the carcinoma class given the dangers of misdiagnosis.

For this classification task, the Inception-ResNet-v2 was used by Ferreria et al. [20]. The classification layers of the base model were replaced by a global average pooling layer, a dense (or fully connected) layer with 256 neurons, a dropout layer with a dropout rate of 0.5, and a final dense layer of 4 neurons. Moreover, the input size of the network was changed to 244×244. Reshaping the images does not significantly impact the form of the cellular structures; however, it does reduce computational cost [20]. The authors did not incorporate stain normalization into their experiments. Data augmentation techniques such as image flips (horizontal and vertical), a 10% zoom range, and shifts (horizontal and vertical) were used to increase the dataset. These particular techniques were chosen with care because if the augmentation causes too much distortion, the anatomical structures in the image could be destroyed [20], and this may result in the network having difficulty extracting discriminative features during training.

Two forms of transfer learning were used in this experiment. At first, only the dense (fully connected) layers of the model were trained. This technique is referred to as feature extraction since the network is using pretrained features (from ImageNet) to classify the breast cancer histology images. The result of this step is that only the weights of the dense layers were adjusted. This aids in overfitting [20]. Afterwards, a certain number of layers were unfrozen so that the network could be fine-tuned. Early stopping with a patience of 20 epochs, and a checkpoint callback monitoring minimum validation loss were the additional techniques implemented to avoid overfitting. The dataset was randomly split into 70% training, 20% validation, and 10% testing. The test set achieved an accuracy of 90%.

In a study by Kassani et al. [21], five different architectures (Inception-v3, Inception-ResNet-v2, Xception, VGG16, and VGG19) were investigated for the classification of the ICIAR2018 dataset. Two stain normalization methods were observed in this study: Macenko et al. [19] and Reinhard et al. [22]. Data augmentation included vertical

flips, contrast adjustment, rotation, and brightness correction. The data was split into 75% and 25% for training and testing, respectively. The images were resized to 512×512 pixels with the help of bicubic interpolation. For each of the models, features were extracted from specific blocks, particularly the layer after a max-pooling layer. The extracted features were put through a global average pooling layer and then concatenated to form a feature vector which was fed into an MLP (multilayer perceptron) set with 256 neurons for final classification. Of these models, the modified Xception network trained with Reinhard stain-normalized images performed the best, with a reported accuracy score of 94%. Overall, the Xception architecture performed the best for both of the stain normalization methods, and the Reinhard [22] technique produced higher accuracies than Macenko [19]. The other architectures ranked in the following order: Inception-v3, Inception-ResNet-v2, VGG16, and VGG19. Interestingly, the approximate parameters for these architectures are 23 million, 54 million, 138 million, and 143 million, respectively. One could hypothesize that an increase in parameter count translates to a decrease in accuracy of this dataset. This indicates that the bigger architectures may have more difficulty extracting critical features from training images, even if measures are taken to enlarge the dataset being used. The results of this study also emphasize the benefit of incorporating stain normalization into preprocessing and how choosing the correct method improves accuracy significantly. Table 1 shows a comparison summary of related deep learning techniques in the literature.

3. Methods and Techniques

The process followed in this research is depicted in Figure 2. The method consisted of two major phases. In the first phase, all the images in the ICIAR2018 dataset were stain-normalized using two techniques: Reinhard [22] and Macenko [19]. For a better understanding of the impact of stain normalization, experiments with nonnormalized images were also conducted. Then specific data augmentation techniques were randomly applied to the images. These augmentation techniques were chosen so that the images would not be too distorted, to avoid the risk of losing distinguishing features. For the second phase, the EfficientNet models were extended to perform classification. For this architecture, sufficient regularization was necessary as the dataset used was relatively small compared to what deep learning models require. This limitation introduces the possibility of overfitting, and employing regularization techniques counteracts this. Figure 3 depicts the proposed method.

3.1. Dataset. The dataset used in this research is called the ICIAR2018 breast cancer histology images dataset [14] and is an extension of the 2015 Bioimaging breast cancer histology images dataset [5]. It contains 400 high-resolution microscopy images and is separated into four classes: normal, benign, *in situ* carcinoma, and invasive carcinoma. All four classes are equally represented. Two medical professionals annotated each image, and if the professionals disagreed on a particular image's annotation, the image was either discarded or confirmed through immunohistochemical analysis. The dataset is available in RGB.tiff format, and each image is 2048×1536 pixels in size, with a pixel scale of $0.42\mu\text{m} \times 0.42\mu\text{m}$ (which refers to the area of tissue covered by a pixel) with a magnification of 200×.

3.2. Preprocessing. Preprocessing is crucial for the classification of histology images. The images in the dataset [14] are rather large while convolutional neural networks are typically designed to take in much smaller inputs. Therefore, the resolution of the images must be decreased so that the network is able to receive the input while maintaining the important features. The size of the dataset is much smaller than what is generally required to train a deep learning model properly; data augmentation is utilized to increase the amount of unique data in the set. This technique contributes toward avoiding overfitting, a phenomenon whereby the model learns the training data well but is entirely unable to generalize and classify unseen images.

3.2.1. Stain Normalization. Many factors contribute to color inconsistencies in histology images, but they are primarily due to the tissue preparation and histology staining process. Other factors may include the conditions and small differences in the labs where the slides are prepared. The techniques used in the process and fixation delays as well as the conditions during slide digitization using a scanner, such as changes in light sources, detectors, or optics, contribute to the discrepancies [4]. These discrepancies in colors in the images could negatively impact the training process in CNNS. [23]. There have been many stain normalization techniques proposed. In this research, two techniques were applied, proposed by Reinhard et al. [22] and Macenko et al. [19].

These techniques aid in improving the efficiency and accuracy of a network by reducing the color inconsistencies in the images. Moreover, without stain normalization, the network may learn staining patterns instead of extracting the relevant features [14].

However, the majority of the top performing methods reported in the "ICIAR2018 Grand Challenge" paper [14] did not use any form of stain normalization, so we also conducted experiments with images that were not normalized.

Images must be converted from the BGR color space to the RGB color space in order for the stain normalization techniques to function as expected.

(1) Macenko Stain Normalization. This technique [19] accounts for the staining protocol used during the preparation of the tissue slide. Firstly, the colors are converted to optical density (OD) via the simple logarithmic transformation.

A value, β, is specified and used as a threshold to remove data with higher OD intensity. Singular value decomposition (SVD) is applied to the optical density tuples from the first step in order to determine a plane. This plane corresponds to

Table 1: Summary of related techniques in the literature.

Reference	Dataset	Pretrained	Architecture	Input size	Stain normalization	Image-wise accuracy
Araújo et al. [5]	Bioimaging 2015	No	Custom CNN	512 × 512	Macenko	4-class: 77.8% 2-class: 80.6%
Vo et al. [17]	Bioimaging 2015	No	3 × Inception-ResNet-v2	600 × 600 450 × 450 300 × 300	Macenko	4-class: 96.4% 2-class: 99.5%
Nawaz et al. [3]	ICIAR2018	Yes	AlexNet	512 × 512	Macenko	81.25%
Ferreria et al. [20]	ICIAR2018	Yes	Inception-ResNet-v2	244 × 244	Nonnormalized	90%
Kassani et al. [21]	ICIAR2018	Yes	VGG16	512 × 512	Macenko Reinhard	83% 87%
Kassani et al. [21]	ICIAR2018	Yes	VGG19	512 × 512	Macenko Reinhard	80% 84%
Kassani et al. [21]	ICIAR2018	Yes	Inception-ResNet-v2	512 × 512	Macenko Reinhard	90% 88%
Kassani et al. [21]	ICIAR2018	Yes	Xception	512 × 512	Macenko Reinhard	91% 94%
Kassani et al. [21]	ICIAR2018	Yes	Inception-v3	512 × 512	Macenko Reinhard	90% 90%

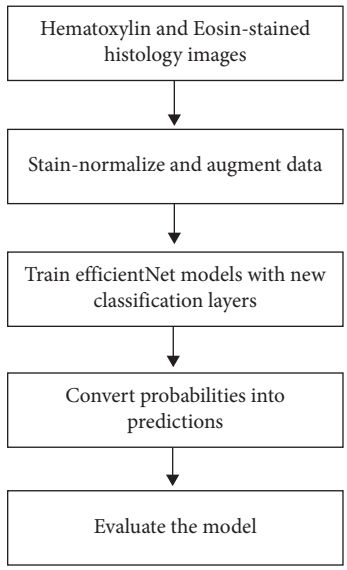


Figure 2: Training process of the experiments.

the two largest singular values found. The optical density-transformed pixels are then projected onto this plane so that the angle at every point concerning the first SVD direction can be determined. Then, the color space transform resulting from the previous steps is applied to the original breast cancer histology image, and the histogram of the image is stretched such that the range covers the lower (100–α)% of the data. Minimum and maximum vectors are calculated and projected back into the optical density space. The hematoxylin stain corresponds to the former vector, and the eosin stain corresponds to the latter vector. The concentrations of the stains are appropriately determined, and the resulting matrix represents the RGB channels and OD intensity. The values α and β are recommended to be set to 1 and 0.15, respectively, and are kept the same for these experiments.

(2) Reinhard Stain Normalization. This technique [22] focuses on mapping the color distribution of an over- or under-stained image to a well-stained image. The use of linear transformation from RGB to $l\alpha\beta$ color space by matching mean and standard deviation values of the color channels achieves this. Essentially, the mean color within the selected target image is transferred onto the source image. This method preserves the intensity variation of the original image. This, in turn, preserves its structure, while its contrast is adjusted to that of the target. In the $l\alpha\beta$ color space, the stains are not precisely separated. The $l\alpha\beta$ color space must be converted back into RGB to attain the normalized image.

Figure 4 shows examples of the stain normalization techniques applied in this study. In this figure, (a) represents the target image that was used for both techniques. Essentially, the techniques aim to normalize the colors in the original images to those of the target. An example of the original image is shown in (b). The subfigures (c) and (d) show the result of using (a) on (b) with the Macenko and Reinhard techniques, respectively.

3.2.2. Data Augmentation. For this step, combinations of methods that are provided by the Keras library were tested to observe the impact on overfitting and contribution to improving classification accuracy. The process of analyzing histology images is rotationally invariant, which means that, irrespective of the angle at which a pathologist views a microscopy image, he/she is still able to examine the image. Therefore, applying a rotation augmentation to the image should not negatively impact the training of the architecture. The rotation augmentation is customized (as in the approach in [5]) such that an image is rotated $(90k)^{\circ}$ in a clockwise direction where $k = \{0, 1, 2, 3\}$. Additionally, a width and height shift, zoom range, and horizontal and vertical flips were randomly applied to rescaled images. This step is implemented in a way that allows the augmentation to be done

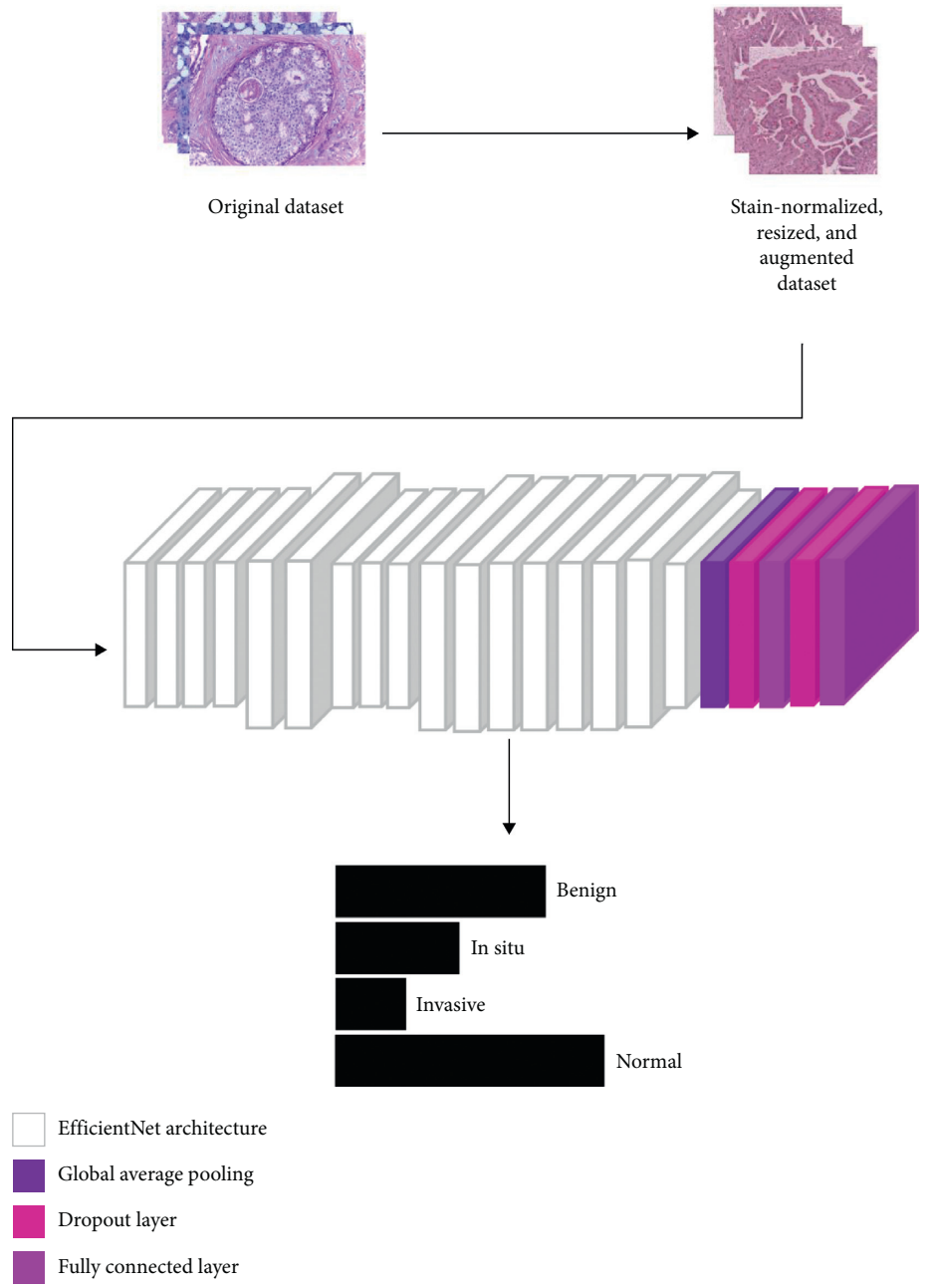
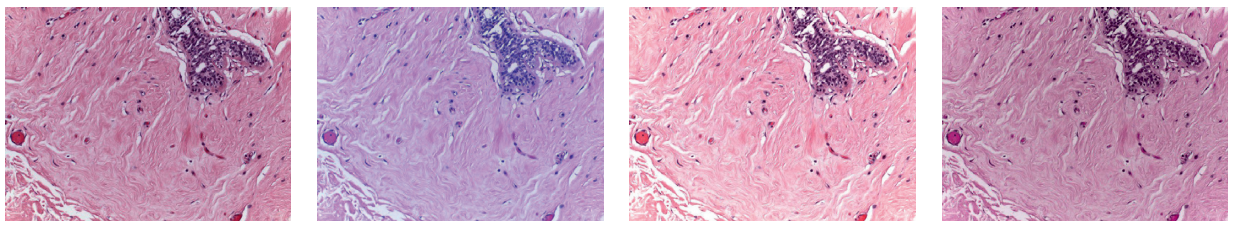


Figure 3: Proposed model for classification of histology microscopy images using deep learning.

dynamically; therefore, no extra storage is required. The normalized images are randomly augmented as they are fed into the model for training. Figure 5 shows an image normalized with Reinhard [22], resized, and randomly augmented with the methods mentioned in Table 2. The images were resized according to the recommended input size of each EfficientNet architecture as shown in Table 3. Table 3 contains the relevant information for each EfficientNet resolution. Resizing the images directly causes a loss of local features but preserves global features in the image. Therefore, the success of the experiments depend on the architecture's ability to recognize and learn these global features [23].

3.3. Transfer Learning. Transfer learning (TL) can be described as follows: "given a source domain $\mathcal{D}_{\mathcal{S}}$ and learning task $\mathcal{T}_{\mathcal{S}}$, a target domain $\mathcal{D}_{\mathcal{T}}$ and learning task $\mathcal{T}_{\mathcal{T}}$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_{\mathcal{T}}$ using the knowledge in $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{T}_{\mathcal{S}}$, where $\mathcal{D}_{\mathcal{S}} \neq \mathcal{D}_{\mathcal{T}}$, or $\mathcal{T}_{\mathcal{S}} \neq \mathcal{T}_{\mathcal{T}}$" [24].

The earlier and middle layers of a CNN detect edges and generic shapes, while the layers toward the end of a CNN detect problem-specific features. The concept of transfer learning is based on utilizing the general features learned in the earlier layers from the source dataset, and a specified number of layers at the end of the model are retrained on the target dataset. The main benefits of TL are saving training



(a) (b) (c) (d)

Figure 4: Examples of original and normalized images: (a) target image; (b) original image; (c) Macenko-normalized; (d) Reinhard-normalized.

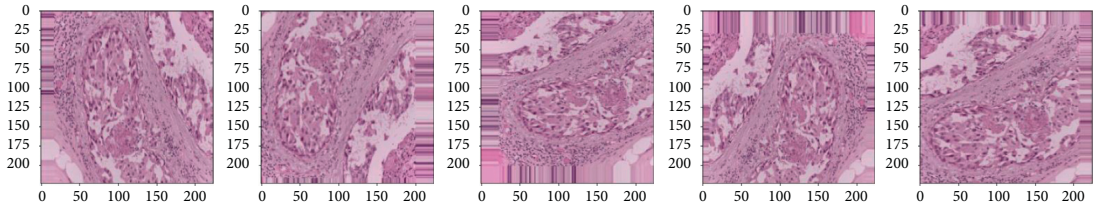


Figure 5: Random data augmentation applied to a normalized image.

Table 2: Values for data augmentation applied to the stain-normalized images.

Augmentation type	Value
Rescaling	1./255
Rotation range	5°
Width shift range	0.1
Height shift range	0.1
Zoom range	0.3
Horizontal flip	True
Vertical flip	True
Additional rotation	0°, 90°, 180°, 270°

Table 3: The number of parameters in each EfficientNet and the recommended input size.

Models	Trainable parameters (million)	Input size
B0	~4.3	224 × 224
B1	~6.8	240 × 240
B2	~8.0	260 × 260
B3	~11.0	300 × 300
B4	~17.9	380 × 380
B5	~28.7	456 × 456
B6	~41.1	528 × 528

time, improving performance of the neural network, and circumventing the limitations caused by lack of data [25]. This technique has been effective in overcoming the issue of small datasets [26].

In the work of Shallu et al. [27], the application of transfer learning for breast cancer histology image classification was investigated. The pretrained networks used in this study were VGG16, VGG19, and ResNet-50. In order to evaluate the effect of using pretrained weights with these models, the authors used the networks as feature generators, extracted features from the images, and used these features to train a logistic regression classifier. The results of these tests were then compared to the results of a full-trained network (trained from scratch with randomly initialized weights). It was proven that fine-tuning significantly impacted the reported precision, recall, F1, accuracy, AUC, and APS scores. The VGG16 model, which is the smallest model investigated (depth-wise, having only 16 convolutional layers) in the experiment, performed the best on the fine-tuning tests, reaching an accuracy of 92.6%. A supposition one can make from the reported fine-tuning results of this study is that the CNN architectures that were larger (depth-wise) had lower accuracy scores: VGG16 obtained 92.6% accuracy, VGG19 obtained 90% accuracy, and ResNet-50 obtained 79.4% accuracy with a train-test split of 90%–10%. This indicates that network capacity is an important factor to consider when choosing a network to fine-tune. A conclusion made in this study was that the fine-tuned networks were more robust to the different sizes of train-test splits than the fully trained networks were.

There are, of course, various challenges with the application of transfer learning to medical image classification, as reported in [28]. One challenge is that medical image classification tasks do not have a sufficient amount of annotated data that is available for training CNNs [29]. This can be attributed to the expense and complexity of the process of annotating images [14]. This lack of data means that large CNNs that generally perform well in applications such as ImageNet would have difficulty avoiding overfitting on these datasets. Therefore, an ample amount of regularization in different forms is needed. Overparameterization is another one of these challenges, and it refers to the great number of parameters in a network. The more trainable parameters a network has, the longer the network will require training, the larger the number of required epochs will be, and the more computation it will require. This is not ideal in the real-world application of these models. A possible way to circumvent these challenges is to use lightweight

architectures that are smaller in size and have fewer parameters, which results in more efficient use of computational power [28]. EfficientNet [13], SqueezeNet [30], and MobileNet-v2 [31] are a few of the recently proposed lightweight architectures.

Multiple forms of transfer learning have been proposed. This includes weight initialization, feature extraction, and fine-tuning. For this application, empirical observations revealed that the combination of feature extraction and fine-tuning did not enhance accuracy. On the contrary, the feature extraction phase could not effectively transfer features from the source dataset to classify the breast cancer histology images. This outcome can be attributed to the fact that the source dataset consists of natural images which bear no resemblance to the histology images. Therefore, high-level features found in the pretrained model's upper layers do not contribute to this specific classification task. These experiments resulted in extreme overfitting even though the various preventative measures were taken. Hence, we can conclude that fine-tuning the architecture with the dataset and utilizing the source dataset's low-level features yield far more acceptable results in this study.

Fine-tuning is described as freezing a certain number of layers in the model such that the generic features extracted at the beginning layers are well utilized. For this study, these generic features come from training on the ImageNet dataset [3, 20]. This dataset contains approximately 14 million natural images with 22 thousand visual categories. Figure 6 depicts the process of fine-tuning.

Choosing the most suitable layer to begin fine-tuning from requires extensive testing. Studies such as [12] have investigated which block to tune from such that results are optimal. The authors [12] concluded that fine-tuning the top layers of a network is much more beneficial than the entire network. However, for this study, fine-tuning began at the third block. We note that beginning fine-tuning from higher blocks results in slight overfitting.

In addition to ImageNet weights, noisy-student weights [32] were also employed to train the models. Empirical observations showed that ImageNet weights were more appropriate for this application.

3.4. EfficientNet. The process of scaling convolutional neural networks is not well understood and is sometimes done arbitrarily until a satisfactory result is found. This process can be tedious because manual tweaking of the relevant parameters is required [13]. The earlier proposed methods of scaling a network include scaling a model by depth [33], by width [34], and by image resolution [35]. Tan and Quoc [13] studied the influence of these scaling methods in a bid to develop a more systematic way of scaling network architecture. The key findings of their research can be summarized into two specific notes: Firstly, scaling up any single dimension of network resolution, depth, or width will improve accuracy; however, this accuracy gain will diminish for larger models. Secondly, to achieve improved accuracy and efficiency, it is essential to balance the dimensions of a network's depth, width, and resolution, instead of focusing on just one of these. Considering these findings, the authors presented a novel scaling method that uses a robust compound coefficient, ϕ, to scale up the networks in a much more structured manner. Equation (1) represents how the authors [13] suggest scaling the depth, width, and resolution with respect to ϕ.

$$\begin{aligned} d &= \alpha^{\phi}, \\ w &= \beta^{\phi}, \\ r &= \gamma^{\phi}, \\ s.t. \quad \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2, \\ \alpha \geq 1, \beta \geq 1, \gamma &\geq 1, \end{aligned} \tag{1}$$

where d, w, and r represent the depth, width, and resolution of the network, respectively, while the constant terms α, β, and γ are determined by a hyperparameter tuning technique called grid search. The coefficient ϕ is user-specified and manages the resources that are available for model scaling. The constants define how the additional resources are assigned to the dimensions in the network. The "floating point operations per second" (FLOPS) is a measure of computer performance [36] and essentially measures how many operations are required to execute the network. If the network's depth is doubled, the number of FLOPS required is doubled too. If the network's width or resolution is doubled, the number of FLOPS required is quadrupled. Therefore, the constraint in (1) indicates that, for any increase in the ϕ value, the new number of FLOPS will increase by 2^{ϕ}. Furthermore, the constant terms must be greater than or equal to one because none of the dimensions should be allowed to be scaled down. The aim of this method [13] is to scale network depth, resolution, and width, such that the accuracy of the network, and the consumption of memory and FLOPS are optimized according to the available resources.

To solidify the concept and prove the effectiveness of the compound scaling method, the authors [13] then developed a mobile-sized baseline network by applying the neural architecture search (a technique used to optimize efficiency and accuracy with respect to FLOPS), which was called the EfficientNet-B0. The model uses inverted residual blocks, consisting of squeeze-and-excitation optimization [37] and swish activation [38]. Swish is defined as

$$\text{Swish}(x) = x * \text{sigmoid}(x). \tag{2}$$

The inverted residual block was introduced in the MobileNet-v2 architecture [31] and makes use of depth-wise separable convolution to decrease the number of parameters and multiplications needed to execute the network. This modification results in faster computation without adversely affecting performance. The inverted block consists of three major components: a convolutional layer (called the expansion layer) which expands the number of channels to prepare the data for the next layer, a depth-wise convolutional layer, and another convolutional layer (the projection

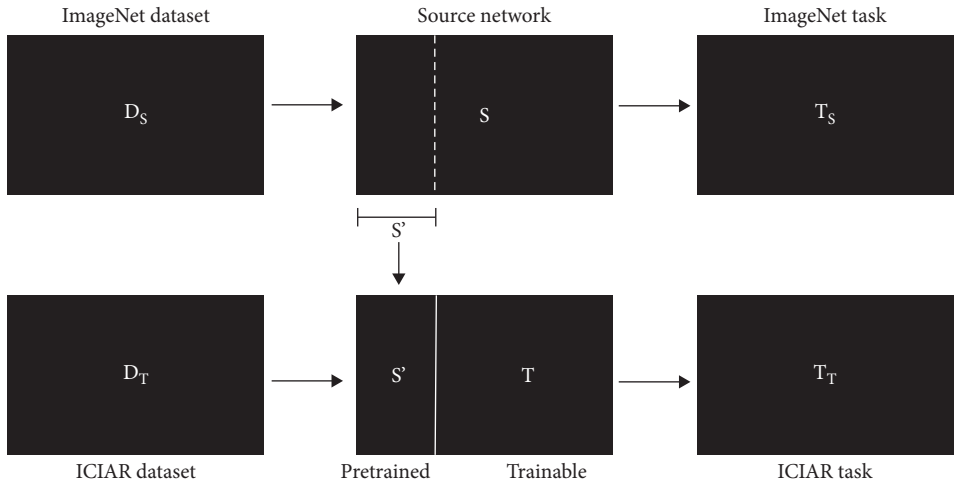


Figure 6: Visualization of fine-tuning.

layer) which is meant to project the data from a large number of channels to a small number of channels. The first and last layers of a residual block are connected via a skip connection. Therefore, during fine-tuning, it is imperative to train entire blocks. Disobeying this restriction can damage the way the network learns [39]. The squeeze-and-excition block consists of a global average pooling (GAP), a reshaping, and two convolutional layers. The GAP layer extracts global features, and then the number of channels is squeezed according to a predefined squeeze ratio.

The compound scaling method was then used to create the EfficientNet family which included the versions B1 to B7; the constants α, β, and γ were fixed; and ϕ was scaled.

The efficacies of the models were tested on the ImageNet dataset and surpassed state-of-the-art convolutional neural networks, with magnitudes being smaller and faster on CPU inference. The outcome (shown in Figure 7) revealed that even though the models have smaller magnitudes than established models in both number of parameters and number of FLOPS, they performed phenomenally.

These models have been successfully used for other histopathology image classification [40–45]. However, at the time of this research, the EfficientNet architecture had not yet been investigated for classification of the ICIAR2018 dataset. We limit our experimentation to the first six EfficientNets due to computational resource restrictions.

3.5. Experimental Settings. In order to ensure that results were reproducible, seeds were set for all packages and methods that allowed this. Specifically, a function was used to split the dataset into the training and validation subsets and was seeded; this value was kept constant throughout all experiments. Therefore, all training and validation images across all experiments are the same.

The dataset was divided into train-test sets with split of 85%–15%, as this split produced the highest results and had less difficulty with overfitting. The images in both subsets were stratified, meaning that the classes were equally represented. Stain normalization of the images was accomplished using a package provided by [46].

The implementation of this approach was dependent on the Keras package within the TensorFlow Python library. Specifically, the Keras ImageDataGenerator [47] was used to create augmentation generators for the training and validation data. Publicly available pretrained EfficientNet models were utilized [48].

Each EfficientNet was extended with a global average pooling layer, a dropout layer with a rate of 0.5, followed by a dense layer of 256 neurons with the ReLU activation function, then another dropout layer with a rate of 0.4, and finally a dense layer of 4 neurons with the softmax activation function. The pooling layer performs subsampling. Essentially, the layer reduces the dimensions of the previous layer by combining the neuron clusters into a single neuron [49]. A dropout rate represents the rate at which input units are set to 0 in a dropout layer. If the units are not set to 0, they are updated and scaled up by a value of 1 – 1/dropoutrate to ensure that the sum over all the inputs does not change. Each neuron in a dense layer is connected to every neuron in the previous layer. Therefore, these layers increase the parameter count in a network substantially. The softmax function enables the output of the network to be a vector of probabilities of the input image belonging to each of the four classes. The Adam optimizer was used with a learning rate of 0.0001; this value was empirically tested and found to produce the best results. When the learning rate is too high, the network learns recklessly and is not able to retain information. When the learning rate is too low, the model training progresses very slowly, which means that more epochs, computational resources, and training time are required. Moreover, early stopping with a patience of 10 epochs, reducing learning rate on plateau with a patience of 8 epochs and minimum learning rate set of 0.000001, and model checkpoints were used during each experiment. Therefore, models with the lowest validation losses were saved while training continued until early stopping ended the execution. The batch size was fixed to 16 for the EfficientNets-B0–B4, the size was set to 5 for the B5 and B6 (due

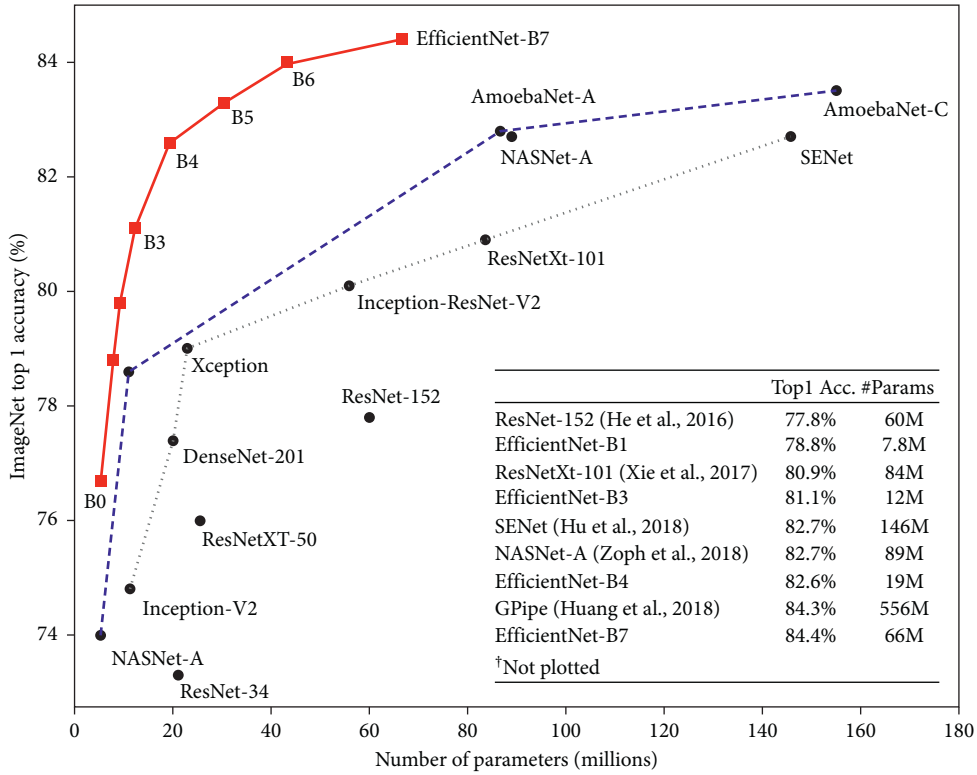


Figure 7: A comparison of EfficientNets with established architectures on the classification of the ImageNet dataset (source: [13]).

to memory constraints), and all models were set to train for 50 epochs. Categorical cross-entropy was employed as the loss function, as it is best suited for multiclass classification. This function is defined as

$$CCE = -\frac{1}{N}\left(\sum_{i=1}^{N} \log P_{model}\left[y_i \in C_{yi}\right]\right), \tag{3}$$

where N represents the number of instances and $\log P_{mo\,de\,l}[y_i \in C_{yi}]$ represents the probability predicted by the model for the i^{th} instance.

All tests were run on the Google Colaboratory platform which provides 25 GB RAM and a 12 GB NVIDIA Tesla K80 graphics processing unit.

4. Results and Discussion

4.1. Evaluation Criteria. The performance of each model was evaluated by calculating the precision, recall, F1-score, and accuracy. The equations below represent the manner in which these metrics are determined.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100, \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100, \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100, \tag{6}$$

$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The precision and recall scores recorded are averaged over all four classes.

4.2. Experimental Results. Table 4 provides the average precision and recall (across the four classes) and the accuracy of each model with the stain normalization methods. Moreover, the average accuracy obtained by the EfficientNet model is calculated.

Accuracy and sensitivity (precision) are of paramount importance in this task. Hence, these metrics will be further analyzed and discussed.

4.3. Discussion

4.3.1. Analysis of Results. The Grand Challenge [14] outcomes revealed that the best performing models were pretrained on ImageNet and fine-tuned. Even though many approaches incorporate patch-wise classification to utilize local features, [3, 5, 14] state that using the entire image for classification was found to produce better results. This insight implies that extracting nuclei and tissue organization features is more valuable for deciphering the image classes than nuclei-scale features [14]. The image-level classification in this research observes the ability of the architecture to

Table 4: Results for the EfficientNet architectures with each stain normalization method.

EfficientNet	Stain normalization technique	Precision	Recall	F1-score	Accuracy
B0	Reinhard	93.65	93.33	93.32%	93.33%
	Macenko	91.67	91.67	91.67%	91.67%
	Nonnormalized	92.37	91.67	91.37%	91.67% 92.9%
B1	Reinhard	94.99	95.00	94.94%	95.00%
	Macenko	94.99	95.00	94.94%	95.00%
	Nonnormalized	93.32	93.33	93.27%	93.33% 94.7%
B2	Reinhard	**98.44**	**98.33**	**98.33%**	**98.33%**
	Macenko	96.88	96.67	96.66%	96.67%
	Nonnormalized	94.03	93.33	93.29%	93.33% **95.3%**
B3	Reinhard	92.50	91.67%	91.62%	91.67%
	Macenko	93.54	93.33	93.27%	93.33%
	Nonnormalized	91.94	91.67	91.66%	91.67% 92.2%
B4	Reinhard	92.38	91.67	91.69%	91.67%
	Macenko	91.94	91.67	91.66%	91.67%
	Nonnormalized	88.56	88.33	88.31%	88.33% 90.6%
B5	Reinhard	88.44	88.33	88.14%	88.33%
	Macenko	91.80	91.67	91.53%	91.67%
	Nonnormalized	92.49	91.67	91.56%	91.67% 90.6%
B6	Reinhard	92.65	91.67	91.67%	91.67%
	Macenko	92.46	91.67	91.68%	91.67%
	Nonnormalized	93.54	93.33	93.39%	93.33% 92.2%

extract global features in the breast cancer histology images and use it to classify unseen images.

The results of the experiments show that the EfficientNet models perform well on the ICIAR2018 dataset. The Reinhard technique [22] outperforms Macenko in the EfficientNet-B0, -B2, and -B4. The Macenko technique [19] returned the highest accuracy for the EfficientNet-B3, and nonnormalized images performed the best for the EfficientNet-B5 and -B6. For the EfficientNet-B1, both Macenko and Reinhard produced the same results. Table 5 shows the average accuracy results obtained by each stain normalization method. In this table, the average results of the nonnormalized method are inferior to that of those Reinhard and Macenko techniques. Therefore, it can be concluded that, on average, applying stain normalization to images as part of the preprocessing step is beneficial.

Notably, the EfficientNet-B2 model produced superior results compared to the other six, which indicates that this architecture was the most successful at extracting and learning the global features in the training set. The average accuracy achieved by this model (95.83%) is higher than those reported in the literature. The additional benefit of this approach is that it is simple, requiring fewer parameters, which implies less training time compared to the previous approaches. The use of images normalized with the Reinhard technique returned the highest sensitivity and accuracy for this model, at 98.33%. Following closely is the result from the use of images stained with the Macenko technique, with a sensitivity and accuracy of 96.67%. The EfficientNet-B1 returns identical results for training the images stained with Reinhard and Macenko, with a sensitivity and accuracy of 95.00%. The difference between the number of parameters in the EfficientNets-B1 and -B2, and the input size both models receive are similar. For this specific dataset, the input sizes of 240×240 and 260×260 are appropriate for successfully extracting global features in the histology images. From Table 4, it is interesting to note that the average accuracy gain seems to increase at first (for EfficientNets-B0–B2), decrease with larger models (for EfficientNets-B3–B5), and then pick up slightly (EfficientNet-B6). The EfficientNet-B0 model has the fewest parameters but does not perform the best. This may be owing to the small input size that the model receives. Resizing the large images to 224×224 may affect the structures in the image, making it more difficult for the model to capture the features. For the larger models that did not perform as well, this may be a consequence of overparameterization relative to the size of the dataset.

Table 6 provides additional insight by showing the models' sensitivity and specificity concerning each class. The *R*, M, and N next to each model name refer to the stain normalization method used on the dataset (Reinhard, Macenko, and nonnormalized). As previously mentioned, high sensitivity is crucial for this problem. It is a widely used metric to determine the proportion of correctly identified

Table 5: The average accuracy results of the stain normalization techniques, Reinhard and Macenko.

Stain normalization technique	Accuracy (avg)
Reinhard	92, 86%
Macenko	93, 10%
Nonnormalized	91, 90%

Table 6: Specificity (spec.) and sensitivity (sens.) of the EfficientNet models.

Model	Normal		Benign		In situ		Invasive	
	Spec. (%)	Sens. (%)	Spec. (%)	Sens. (%)	Spec. (%)	Sens. (%)	Spec. (%)	Sens. (%)
B0 (R)	93.33	93.33	87.0	93.33	100	6.7	93.75	100
B1 (R)	92.86	86.7	93.75	100	3.3	93.3	100	100
B2 (R)	93.5	100	100	100	00	93.3	100	100
B3 (M)	100	6.67	86.67	86.7	93.75	100	93.75	100
B4 (R)	100	6.67	92.86	86.7	83.33	100	3.33	93.3
B5 (N)	100	0.00	92.86	86.7	83.33	100	93.75	100
6 (N)	93.33	93.33	87.50	93.33	93.3	93.33	100	93.33

positives (as shown in (6)). As expected, the EfficientNet-B2 model shows consistent high sensitivity throughout all four classes; however, it is less sensitive toward the *in situ* carcinoma class. The EfficientNet-B4 and -B5 return the lowest sensitivity to all four classes but especially to the noncarcinoma classes. The sensitivity toward the carcinoma class must be maximized, as an incorrect prediction of this tissue could lead to misdiagnosis and severe consequences. Aresta et al. [14] state that the benign class affects the performance of the networks due to the structural similarity with normal tissue. This class contains more morphological variability than the others, which translates to increased difficulty in learning discriminative features. This statement is consistent with the results reported in this study, as the benign class typically returns the lowest sensitivity relative to the other classes.

4.3.2. Analysis of Accuracy and Loss Curves. The graphs in Figure 8 show the accuracy and loss curves for three models: the EfficientNet-B2, -B4, and -B6 trained with Reinhard-normalized images. The curves represent the progress of the training and validation accuracy and loss during the training of the models. These metrics are recorded per epoch.

In graphs (a) and (b), both curves are very close to each other. The evident noise in the curves is caused by the data augmentation, as no single image is fed into the convolutional neural network twice. In graphs (c) and (d), it is obvious that the training is proceeding well; however, the validation accuracy and loss are not. This indicates that the B4 architecture is learning features adequately but it is not able to generalize at the same level. Hence, the validation loss curve is extremely noisy and does not fit perfectly. Finally, in graphs (e) and (f), the architecture clearly has difficulty learning the distinguishing features. The training curve in (e) does not increase at the same rate that (a) and (c) do. Similarly, in (f), the validation loss does not descend below 0.2 as it does for the B2 and B4 models in (b) and (d), respectively. A possible reason for the validation loss curve being below the training loss curve in (f) is that the architecture finds the validation set to be unrepresentative of the entire dataset and easier to predict than the training set [52].

Figure 9 shows the confusion matrices corresponding to the graphs in Figure 8. From these matrices, it can be seen that the larger models do not predict the noncarcinoma classes well. This indicates that the larger models do not learn the features from the normal and benign classes well.

4.3.3. Comparison with Previous Approaches. The results of this study must be compared with similar approaches to understand the effect of using the EfficientNets instead of other architectures. Table 7 summarizes the pretrained architectures used in previous approaches and the accuracies obtained for four-class, image-wise classification. It is worth noting that, among the architectures listed in this table, the EfficientNet-B2 has the lowest number of parameters (approximately 8 million) and is computationally cheaper than the other approaches, as it is a lightweight model. The ResNet-50 and Inception-v3 networks also perform well on this dataset, and both architectures have around 23 million parameters. On the other hand, the VGG16 and VGG19 architectures have over 100 million parameters and produce inferior results. This shows that smaller architectures produce higher accuracies for this dataset. The results prove that transfer learning with the EfficientNet-B2 yielded superior results in comparison to other architectures. Both the Reinhard stain normalization technique and Macenko stain normalization technique yielded satisfactory results on the EfficientNet-B2.

4.3.4. Challenges with This Approach. The major challenge faced in these experiments was combatting overfitting. Due to the size of the dataset, overfitting on the large architectures was ensured. Incorporating data augmentation techniques to increase the number of unique samples that the CNN would see was not sufficient, so other forms of regularization had to be explored. These forms include dropout layers, early stopping, and model checkpointing. The appropriate dropout rate was found through a grid search,

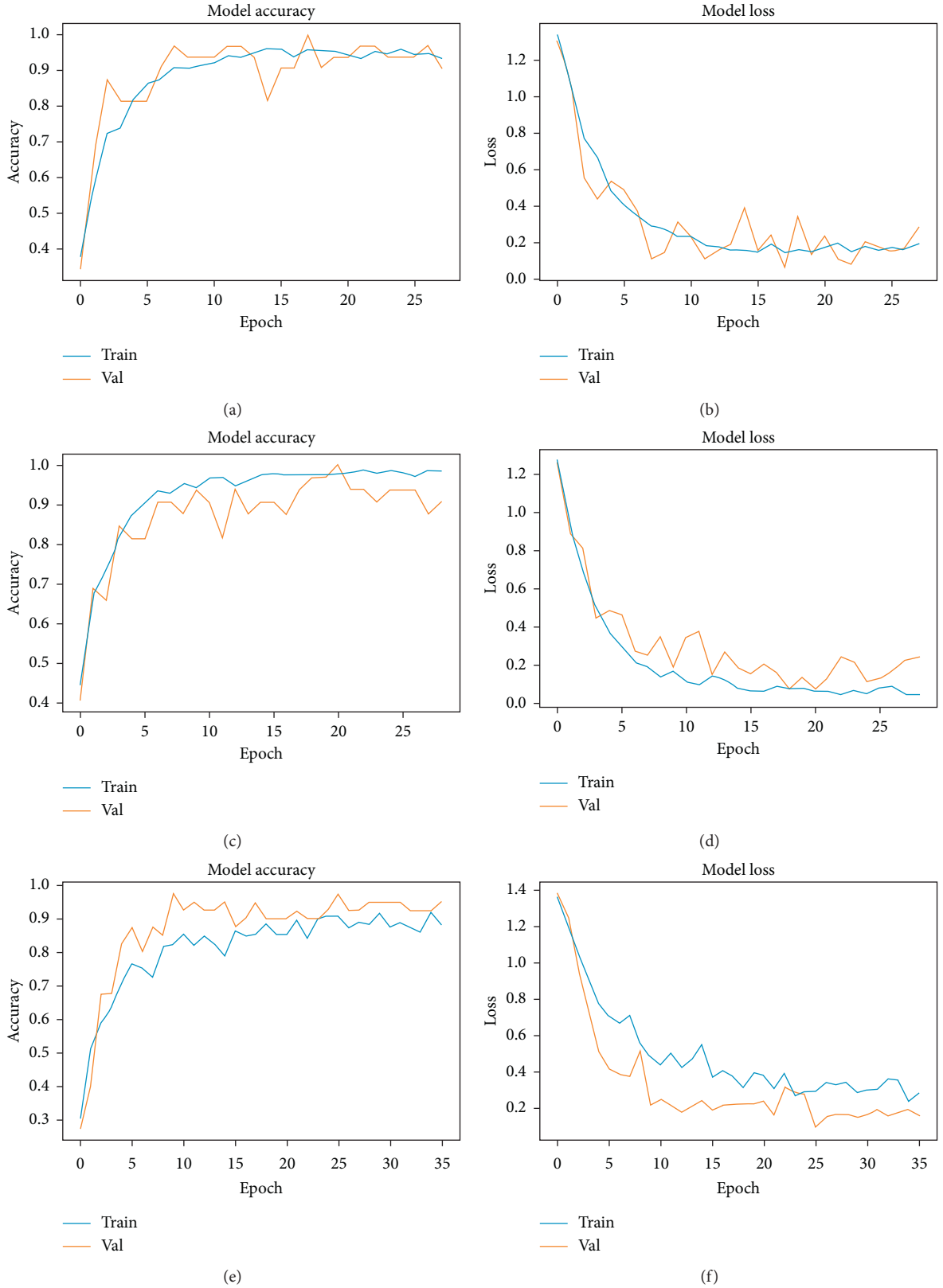


Figure 8: Noteworthy accuracy and loss graphs: (a) EfficientNet-B2 Reinhard accuracy graph; (b) EfficientNet-B2 Reinhard loss graph; (c) EfficientNet-B4 Reinhard accuracy graph; (d) EfficientNet-B4 Reinhard loss graph; (e) EfficientNet-B6 Reinhard accuracy graph; (f) EfficientNet-B6 Reinhard loss graph.

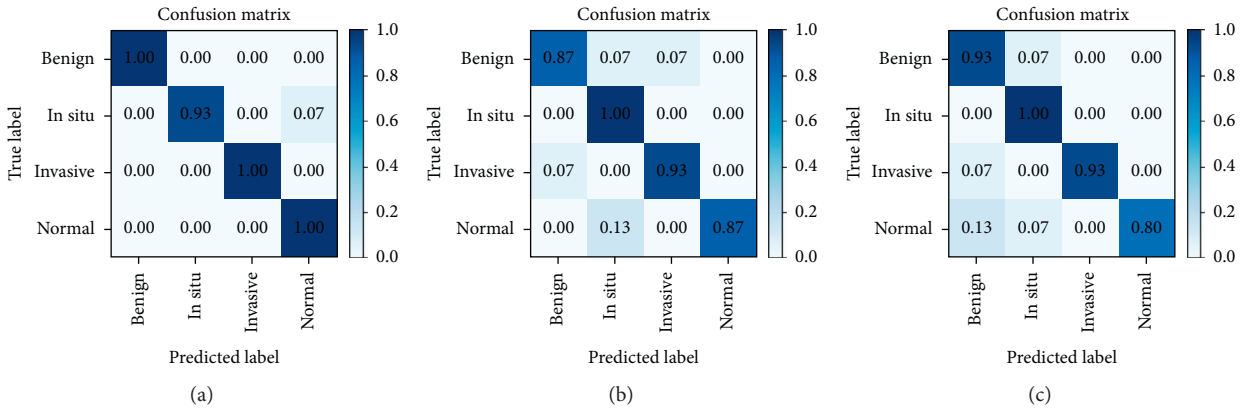


Figure 9: Confusion matrices for the three models: (a) EfficientNet-B2; (b) EfficientNet-B4; (c) EfficientNet-B6.

Table 7: Comparison with previous approaches using pretrained architectures for classification of the ICIAR2018 dataset.

Reference	Architecture	Stain normalization	Accuracy
Nawaz et al. [3]	AlexNet	Macenko	81.25%
Ferreria et al. [20]	Inception-ResNet-v2	None	90%
Kassani et al. [21]	VGG16	Macenko	83%
		Reinhard	87%
Kassani et al. [21]	VGG19	Macenko	80%
		Reinhard	84%
Kassani et al. [21]	Inception-ResNet-v2	Macenko	90%
		Reinhard	88%
Kassani et al. [21]	Xception	Macenko	91%
		Reinhard	94%
Kassani et al. [21]	Inception-v3	Macenko	90%
		Reinhard	90%
Golatkar et al. [50]	Inception-v3	Vahadane	85%
Vesal et al. [51]	Inception-v3	Reinhard	97.08%
Vesal et al. [51]	ResNet-50	Reinhard	96.66%
Our approach	EfficientNet-B2	Reinhard	**98.33%**
Our approach	EfficientNet-B2	Macenko	96.67%

along with optimal batch size and number of dense layers. Alternative loss and activation functions were investigated but contributed more to overfitting and unstable learning. ReLU was chosen as the activation function for this network as it is computationally efficient, is simple, and has been empirically proven to work well [53].

4.3.5. Limitations of the Study. The main objective of this research was to observe the ability of EfficientNet architecture to classify the images of the ICIAR2018 dataset into four classes: normal, benign, invasive carcinoma, and *in situ* carcinoma. However, classifying the images into the groups *carcinoma* and *noncarcinoma* is valuable and helpful pursuit and should return improved accuracies [5, 17]. This binary classification requires the efficient extraction of local features in the images. The architectural additions proposed in this study perform well in extracting global features for the multiclass classification but do not perform equally in the binary classification. A reasonable explanation for this is that the loss of information caused by resizing images translated to the network having difficulty in locating discriminative local features.

5. Conclusion

The EfficientNet family is a set of state-of-the-art convolutional neural networks that achieved preeminent results on established image datasets by carefully balancing the crucial dimensions in a network such that accuracy and efficiency were maximized. The field of medical image analysis suffers from a paucity of publicly available data, which results in researchers having to utilize small and, often, unbalanced datasets. Transfer learning techniques offer support in this area by enabling the reuse of generic features from large source datasets for small target datasets.

In this research, the application of seven versions of EfficientNets with transfer learning for breast cancer histology image classification was investigated. Transfer learning was employed in the form of fine-tuning. The experimental results confirm that this architecture was able to effectively extract and learn the global features in an image, such as the tissue and nuclei organization. These features were then used to classify an image into four classes: normal, benign, invasive carcinoma, and *in situ* carcinoma. Among the seven models tested, the EfficientNet-B2 architecture which has approximately 8 million parameters produced superior results with an accuracy of 98.33% and sensitivity of 98.44%. The results achieved showed that the number of feature maps and the number of parameters (8.0 million) for EfficientNet-B2 are optimum for the research problem in question. Furthermore, the effects of two different stain normalization techniques were observed, and the outcomes were compared to the use of nonnormalized images. From the results of the experiments, no specific stain normalization proved to be superior to the other. Instead, the smaller models performed better with the Reinhard technique, and the larger models performed better with no stain normalization.

For future work, considering the results of this research, it would be interesting to observe the impact of an ensemble

of EfficientNets for this application. Furthermore, the possibility of these lightweight architectures performing well on other histology image datasets such as PatchCamelyon would be worth exploring [54]. Since multistage transfer learning has proved to be advantageous [55], investigating this technique for breast cancer histology images may produce similar enhanced accuracies. Therefore, this is also a path worth pursuing. Lastly, we expect to explore whether the proposed model is suitable for incremental learning.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics 2015," *CA: A Cancer Journal for Clinicians*, vol. 65, pp. 5–29, 2015.

[2] American Cancer Society, *Breast Cancer Facts & Figures 2019-2020"*, 2019, https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf, Accessed 23-June-2020.

[3] W. Nawaz, S. Ahmed, M. Tahir, and H. Khan, "Classification of breast cancer histology images using ALEXNET," in *Proceedings of the16th IEEE International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 869–876, Bhurban, November 2018.

[4] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: a review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[5] T. Araújo, G. Aresta, E. Castro et al., "Classification of breast cancer histology images using convolutional neural networks," *PLOS ONE*, vol. 12, no. 6, pp. 1–14, 2017.

[6] J. G. Elmore, G. M. Longton, P. A. Carney et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA*, vol. 313, no. 11, pp. 1122–1132, 2015.

[7] A. Chan and J. A. Tuszynski, "Automatic prediction of tumour malignancy in breast cancer with fractal dimension," *Royal Society Open Science*, vol. 3, 2016.

[8] B. Maria Priego-Torres, D. Sanchez-Morillo, M. A. Fernandez-Granero, and M. Garcia-Rojo, "Automatic segmentation of whole-slide h&e stained breast histopathology images using a deep convolutional neural network architecture," *Expert Systems With Applications,*, Article ID 113387, 2020.

[9] R. Bhargava and A. Madabhushi, "Emerging themes in image informatics and molecular analysis for digital pathology," *Annual Review of Biomedical Engineering*, vol. 18, no. 1, pp. 387–412, Article ID 27420575, 2016.

[10] S. Z. Ramadan, "Methods used in computer-aided diagnosis for breast cancer detection using mammograms: a review," *Journal of Healthcare Engineering*, vol. 2020, Article ID 9162464, 21 pages, 2020.

[11] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: a review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[12] I. Kandel and M. Castelli, "How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset," *Applied Sciences*, vol. 10, no. 10, p. 3359, 2020.

[13] M. Tan and V. Quoc, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019, https://arxiv.org/abs/1905.11946.

[14] G. Aresta, T. Araújo, S. Kwok et al., "Bach: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019.

[15] M. Kowal, "Computer-aided diagnosis for breast tumor classification using microscopic images of fine needle biopsy," *Advances in Intelligent Systems and Computing*, vol. 230, p. 9, 2013.

[16] A. D. Belsare, M. M. Mushrif, M. A. Pangarkar, and N. Meshram, "Classification of breast cancer histopathology images using texture feature analysis," in *Proceedings of the TENCON 2015-2015 IEEE Region 10 Conference*, pp. 1–5, Macau, China, November 2015.

[17] D. Vo and Q. Nguyen, "Classification of breast cancer histology images using incremental boosting convolution networks," *Information Sciences*, vol. 482, pp. 123–138, 2018.

[18] T. Araújo, G. Aresta, E. Castro et al., *Bioimaging Challenge 2015 Breast Histology Dataset*, 2017, https://rdm.inesctec.pt/dataset/nis-2017-003 Accessed 23-June-2020.

[19] M. Macenko, M. Niethammer, J. S. Marron et al., "A method for normalizing histology slides for quantitiative analysis," in *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, pp. 1107–1110, Boston, MA, USA, July 2009.

[20] C. A. Ferreira, P. S. Tânia Fernandes Melo, M. I. Meyer, E. Shakibapour, P. Costa, and A. Campilho, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," *Image Analysis and Recognition*, pp. 763–770, 2018.

[21] S. H. Kassani and P. Hosseinzadeh Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, Breast cancer diagnosis with transfer learning and global pooling," in *Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, October 2019.

[22] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 4, pp. 34–41, 2001.

[23] Y. Guo, H. Dong, F. Song, C. Zhu, and J. Liu, "Breast cancer histology image classification based on deep neural networks," *Lecture Notes in Computer Science*, pp. 827–836, 2018.

[24] S. Pan and Q. Yang, "A survey on transfer learning. Knowledge and Data Engineering," *IEEE Transactions on*, vol. 22, pp. 1345–1359, 2010.

[25] N. Donges, *What Is Transfer Learning? Exploring the Popular Deep Learning Approach*, 2019, https://www.builtin.com/data-science/transfer-learning Accessed 23-June-2020.

[26] L. Alzubaidi, O. Al-Shamma, M. Fadhel, and L. Farhan, "Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model," *Electronics*, vol. 9, p. 445, 2020.

[27] A. Shallu and R. Mehra, *Breast Cancer Histology Images Classification: Training from Scratch or Transfer Learning?*, 2018, https://www.semanticscholar.org/paper/Breast-cancer-histology-images-classification%3A-from-Shallu-Mehra/6e4e28590f9ba32be410a94cc896560fa69e2db2.

[28] D. Zeng Rajesh Godasu and K. Sutrave, *Transfer Learning in Medical Image Classification: Challenges and Opportunities*, 2020, https://aisel.aisnet.org/mwais2020/18/.

[29] A. Muhammad Dawud, K. Yurtkan, and H. Oztoprak, "Application of deep learning in neuroradiology: brain haemorrhage classification using transfer learning," *Computational Intelligence and Neuroscience*, vol. 2019, Article ID 4629859, 12 pages, 2019.

[30] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," 2016, https://arxiv.org/abs/1602.07360.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," 2018, https://arxiv.org/abs/1801.04381.

[32] M. Innat, *Efficientnet Keras Noisy-Student Weights B0-B7*, 2020, https://www.kaggle.com/ipythonx/efficientnet-keras-noisystudent-weights-b0b7?select=efficientnet-b1_noisy-student_notop.h5.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, https://arxiv.org/abs/1512.03385.

[34] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, https://arxiv.org/abs/1605.07146.

[35] Y. Huang, Y. Cheng, A. Bapna et al., "Gpipe: efficient training of giant neural networks using pipeline parallelism," 2018, https://arxiv.org/abs/1811.06965.

[36] *Wikipedia Contributors. Flops — Wikipedia, the Free Encyclopedia*, [Online; accessed 23-June-2020], 2020.

[37] J. Hu, Li Shen, and G. Sun, "Squeeze-and-excitation networks," 2017, https://arxiv.org/abs/1709.01507.

[38] P. Ramachandran, B. Zoph, and V. L. Quoc, "Searching for activation functions," 2017, https://arxiv.org/abs/1710.05941.

[39] Keras Team, *Keras Documentation: Image Classification via Fine-Tuning with Efficientnet*, 2020, https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/.

[40] Y. Sun, F. Amira Binti Hamzah, and B. Mochizuki, "Optimized light-weight convolutional neural networks for histopathologic cancer detection," in *Proceedings of the 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 11–14, Kyoto, Japan, March 2020.

[41] V. Khobragade, N. Jain, and D. Singh Sisodia, "Deep transfer learning model for automated screening of cervical cancer cells using multi-cell images," *Communications in Computer and Information Science*, pp. 409–419, 2020.

[42] J. Wang, Q. Liu, H. Xie, Z. Yang, and H. Zhou, "Boosted Efficientnet: detection of lymph node metastases in breast cancer using convolutional neural network," 2010, https://arxiv.org/abs/2010.05027.

[43] Q. Ha, Bo Liu, Nvidia, and F. Liu, "Identifying melanoma images using efficientnet ensemble: winning solution to the siim-isic melanoma classification challenge," 2020, https://arxiv.org/abs/2010.05351.

[44] C. Mohamed and A. Moulay, "Akhloufi. Explainable diabetic retinopathy using efficientnet∗," in *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, Canada, July 2020.

[45] Y. J. Suh, J. Jung, and B.-J. Cho, "Automated breast cancer detection in digital mammograms of various densities via deep learning," *Journal of Personalized Medicine*, vol. 10, no. 4, p. 211, 2020.

[46] P. Byfield, *Peter554/staintools*, 2020, https://github.com/Peter554/StainTools.

[47] Keras Team, *Keras Documentation: Image Data Preprocessing*, 2020.

[48] P. Yakubovskiy, *Qubvel/efficientnet*, 2020, https://github.com/qubvel/efficientnet.

[49] S. Kalvankar, H. Pandit, P. Parwate, and K. Wagh, "Galaxy morphology classification using efficientnet architectures," *MNRAS*, pp. 1–12, 2020.

[50] A. Golatkar, D. Anand, and S. Amit, "Classification of breast cancer histology using deep learning," 2018, https://arxiv.org/abs/1802.08080.

[51] S. Vesal, N. Ravikumar, A. Abbas Davari, S. Ellmann, and A. Maier, "Classification of breast cancer histology images using transfer learning," 2018, https://arxiv.org/abs/1802.09424.

[52] J. Brownlee, *How to Use Learning Curves to Diagnose Machine Learning Model Performance*, 2019, https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/.

[53] *Wikipedia Contributors. Rectifier (Neural Networks) — Wikipedia, the Free Encyclopedia*, Online; accessed 5-January-2021, 2020.

[54] S. V. Bastiaan, J. Linmans, Jim Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," 2018, https://arxiv.org/abs/1802.09424.

[55] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, and K. H. Cha, "Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE Transactions on Medical Imaging*, vol. 38, no. 3, pp. 686–696, 2019.

Hindawi
Computational Intelligence and Neuroscience
Volume 2020, Article ID 8826568, 16 pages
https://doi.org/10.1155/2020/8826568

Research Article

A Semisupervised Learning Scheme with Self-Paced Learning for Classifying Breast Cancer Histopathological Images

Sarpong Kwadwo Asare,[1] **Fei You**,[1] **and Obed Tettey Nartey**[2]

[1]*School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[2]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China*

Correspondence should be addressed to Sarpong Kwadwo Asare; sk_asare@std.uestc.edu.cn

Received 25 September 2020; Revised 2 November 2020; Accepted 5 November 2020; Published 8 December 2020

Academic Editor: Vahid Rakhshan

The unavailability of large amounts of well-labeled data poses a significant challenge in many medical imaging tasks. Even in the likelihood of having access to sufficient data, the process of accurately labeling the data is an arduous and time-consuming one, requiring expertise skills. Again, the issue of unbalanced data further compounds the abovementioned problems and presents a considerable challenge for many machine learning algorithms. In lieu of this, the ability to develop algorithms that can exploit large amounts of unlabeled data together with a small amount of labeled data, while demonstrating robustness to data imbalance, can offer promising prospects in building highly efficient classifiers. This work proposes a semisupervised learning method that integrates self-training and self-paced learning to generate and select pseudolabeled samples for classifying breast cancer histopathological images. A novel pseudolabel generation and selection algorithm is introduced in the learning scheme to generate and select highly confident pseudolabeled samples from both well-represented classes to less-represented classes. Such a learning approach improves the performance by jointly learning a model and optimizing the generation of pseudolabels on unlabeled-target data to augment the training data and retraining the model with the generated labels. A class balancing framework that normalizes the class-wise confidence scores is also proposed to prevent the model from ignoring samples from less represented classes (hard-to-learn samples), hence effectively handling the issue of data imbalance. Extensive experimental evaluation of the proposed method on the BreakHis dataset demonstrates the effectiveness of the proposed method.

1. Introduction

Breast cancer is one of the most frequent cancers among women and the second most common cancer globally, affecting about 2.1 million women yearly. Statistics from a global cancer report recorded that an estimated 627,000 women died from breast cancer in 2018 [1]. This figure is approximately 15% of all cancer deaths among women. Also, a recent report from the American Cancer Society's forecast for 2019 predicts that there will be almost 286,600 new cases of invasive breast cancer, about 63,930 new noninvasive cases, and about 41,760 deaths among women in the United States [2]. This worrisome trend necessitates the need for automated breast cancer detection and diagnosis [3]. Computer-aided detection or diagnosis (CAD) systems can contribute significantly in the early detection of breast cancer. Early detection is vital as it can help in reducing the morbidity rates among breast cancer patients [4].

Existing manual methods for breast cancer diagnosis include the use of radiology images in identifying areas of abnormalities. These images, however, cannot be used to accurately determine cancerous areas [5]. Biopsy [6] does help to identify a cancerous area in an image. Breast tissue biopsies help pathologists to histologically assess the microscopic structure and elements of breast tissues. The outcome of biopsy still requires a histopathologist to double-check on the results since a confirmation from a histopathologist is the only clinically accepted method. However, since the diagnosis provided by biopsy tissue and hematoxylin and eosin stained images is nontrivial, there is often some disagreements on the final diagnosis by histopathologists [7]. The drawbacks associated with the methods

mentioned above drive the need for computer-aided systems for breast cancer diagnosis systems to improve diagnosis efficiency, increase the diagnosis concordance between specialists, reduce time, and lessen the burden on histopathologists [4, 8].

Deep convolutional neural networks (CNNs) have achieved tremendous successes in several disciplines including but not limited to object detection [9, 10], segmentation [11], and classification [12, 13]. Recent advancements in machine learning and deep learning in medical diagnosis are motivating lots of research in the classification of breast cancer histopathological images [14, 15]. The build nature of CNNs makes them capable of learning hierarchical feature representation from categorical data, and this is the underlying principle behind the success of CNNs in accomplishing tasks. In the specific case of breast cancer classification, existing work in the literature has adopted CNNs in achieving state-of-the-art results. Some of these methods mentioned in the literature are based on hand-engineered features [16–18]. However, methods that rely on hand-crafted features are inefficient and not robust, and they merely extract sufficient features that are beneficial in classifying histopathological images, not to mention that the entire process is a laborious and computationally expensive one. Other methods mentioned in the literature adopt deep learning approaches for classifying breast cancer histopathological images. Deep learning methods offer a better alternative to methods that rely on hand-engineered features, achieving excellent performances in many classification tasks [19–22]. Convolutional neural networks in particular have achieved state-of-the-art performances in classifying breast cancer histopathological images. In [23], the authors compared two machine learning schemes for binary and multiclass classification of breast cancer histology images. In the first approach, the authors extracted a set of hand-crafted features via bag of words and locality-constrained linear coding. They trained these features with support vector machines. Next, they experimented with a combination of hand-engineered features with a CNN as well as CNN features with the classifier's configuration. On the BreakHis dataset, the authors reported accuracy between 96.15% and 98.33% for binary classification and accuracy between 83.31% and 88.23% for multiclassification. Similar successes have also been reported in [8, 24, 25].

In spite of these successes, it is also pertinent to note that the deep layers associated with CNN models imply the fact that they require large amounts of well-labeled data during training to achieve satisfactory results. Training on relatively small amount of data leaves the models prone to overfitting and, subsequently, poor generalization. In the medical imaging domain, obtaining abundant labels for image samples is a major challenge, not to mention that a large amount of image samples are also required to aid in a model's ability to generalize well on data. Again, the process of labeling image samples is a time-consuming and an expensive one, requiring expertise knowledge. Existing methods mentioned in the literature that perform classification of histopathological images resort to training CNN models with random initialization and data augmentation techniques in a bid to improve a model's performance [23, 25, 26]. Such an approach enables a model to adapt to new data patterns on its own with augmented data samples that improve the number of training samples. These methods typically use only labeled data, since the learning process involved is a supervised one. However, an effective way of reducing labeling cost and generating more training samples is to make use of labeled and unlabeled data, via semisupervised learning (SSL) [27, 28]. Semisupervised learning aims to incorporate both labeled and unlabeled data in building better learners by fully considering the supervised knowledge delivered by labeled data and unsupervised data structure under unlabeled ones [27]. At the heart of semisupervised learning is training a learner on labeled data and using the learner to predict labels for unlabeled data. Moreover, compared to the process of obtaining well-labeled data, unlabeled data is rather inexpensive and abundant. Semisupervised learning algorithms have been adopted in some works mentioned in the literature for some classification tasks [27, 29–34].

In [35], the authors reported a cost-effective active learning approach for classifying deep images. Their proposed approach first progressively feeds samples from the unlabeled data into the CNN. Then clearly classified samples and the most informative samples are selected via a selected criterion and applied on the classifier of the CNN. The CNN model is then updated after adding user-annotated minority uncertain samples to the labeled set and pseudolabeling the majority certain samples. However, this approach acquires the least certain unlabeled examples for labeling and while simultaneously assigning predicted pseudolabels to most certain examples, and such a technique is not always helpful [36]. In [30], the authors use both labeled and unlabeled data for training a deep model across learning cycles. The authors employed both unsupervised feature learning and semisupervised learning. Unsupervised feature learning is used on all data once at the beginning of the active learning pipeline and the resulting parameters are used to initialize the model at each active learning cycle. The authors used semisupervised learning on all data at every learning cycle, replacing supervised learning on labeled examples alone, which is typical of tradition active learning methods. The approach adopted in this work parallels the works in [30, 37] in that a pseudolabel is generated for each unlabeled example but it differs from the work in [37] in that all unlabeled ones are pseudolabeled as opposed to only the majority high-confidence samples. This work employs semisupervised learning with self-training for training a classifier, rather than employing active learning. The work in [29] tackles the issue of classical multimedia annotation problems ignoring the correlations between different labels by combining label correlation mining and semisupervised feature selection into a single framework. Their approach utilizes both labeled and unlabeled data to select features while label correlations and feature corrections are simultaneously mined. In contrast, unlike selecting features via semisupervised learning, our work generates pseudolabels for the unlabeled samples and selects the most confident pseudolabeled samples via the pseudolabel generation and selection algorithm. By incorporating the self-paced learning concept into the selection

process, the model learns samples from both well- and less-represented classes, which tackles the issue of model bias when selecting samples. The base model then learns features from both the labeled data and the selected pseudolabeled samples during training. We also solve the issue of class imbalance by introducing a class balancing framework. These two issues were not addressed in their work.

In [31], the authors proposed a semisupervised model named adaptive semisupervised feature selection for cross modal retrieval. In their semisupervised framework, the labels for unlabeled data are predicted by the graph-based label propagation. Then the unlabeled data with the predicted labels are combined with the labeled data to learn the mapping matrices. Meanwhile, the mapping matrices update the predicted label matrices, which can ensure that the raw feature distribution will be as consistent as possible with the semantic distribution in the subspace after several iterations. Our work parallels this proposed work with respect to predicting labels for unlabeled data and combining both the predicted labels with labeled data in updating training data for another iterative. The differences lie in the fact that our approach first uses the base learner to predict pseudolabels for the unlabeled samples after first training the learner with labeled samples, rather than graph-based label propagation. Then, a pseudolabel selection algorithm selects the most confident pseudolabeled sampled samples before updating the training samples with these selected pseudolabeled samples and labeled samples via self-training. This contrasts mapping matrices which are used to update the predicted label matrices in their approach. Again, our work focuses on generating confident pseudolabeled samples to augment the training data, making more reliable data available to the learner during training, as well as solving the issue of class imbalance in the data set while ensuring the fact that the model exhibits fairness in the selection process by learning from both well- and less-represented samples. Also, the work in [32] introduces a novel discriminative least squares regression (LSR) which equips each label with an adjustment vector. This technique avoids incorrect penalization on samples that are far from the boundary and at the same time facilitates multiclass classification by enlarging the geometrical distance of instances belonging to different classes. The authors assign a probabilistic vector fit each sample, hence ensuring the importance of labeled data while characterizing the contribution of unlabeled instance according to its uncertainty. Our approach primarily focuses on the generation of reliable pseudolabeled samples in augmenting the training data. The reliability of a pseudolabeled sample is determined by the pseudolabel selection algorithm which ensures the selection of pseudolabeled samples with the most confident probability. This prevents the situation where incorrectly labeled samples are added to the training samples. Also, our semisupervised learning approach hinges on the concept self-training and self-paced learning, which distinguishes our approach from the one reported in our work. The similarities lie in the fact that their proposed work and ours utilize both labeled and unlabeled data in the learning process.

To this end, this work proposes a novel semisupervised learning framework that uses self-training and self-paced learning (SPL) [38] to classify breast cancer histopathological images. Self-training is a semisupervised technique capable of learning a better decision boundary for labeled and unlabeled data. Self-training is accomplished by alternating between the generation of a set of pseudolabels corresponding to a large selection scores in the unlabeled-target domain and training a network (usually by fine-tuning) based on these selected pseudolabels and their corresponding pseudolabeled samples and labeled training data. The assumption here is that the target samples with higher prediction probability are right and have better prediction accuracy. In the proposed method, the process of generating and selecting pseudolabels is achieved via a novel pseudolabel generation and selection algorithm that selects only pseudolabels with the highest probability. The selection process is based on SPL, where in the initial learning stage, "easy" samples are selected and then "hard-to-transfer" samples are gradually added in a meaningful manner, making the classifier more robust. In a nutshell, the main contributions of this work are as follows:

We propose a novel semisupervised learning framework that utilizes self-training with self-paced learning in classifying breast cancer histopathological images by formulating the problem as a loss minimization scheme which can be solved using an end-to-end approach.

We introduce a novel pseudolabel generation and selection algorithm for selecting pseudolabels with relatively high-confidence probabilities to augment the training samples for retraining the model. In retraining the model, the optimization process begins by selecting pseudolabeled samples with relatively higher confidence ("easy" samples) then gradually adds "hard" samples to the training data. This ensures the selection of pseudolabels with high precision and prevents mistake reinforcement.

To tackle the issue of class imbalance associated with self-training methods when generating and selecting pseudolabels, we implement confidence scores that use class-wise normalization in generating and selecting pseudolabels with balanced distribution.

We obtain significant accuracy performance on the BreakHis dataset compared to the state-of-the-art approaches.

2. Methods

We provide an overview of the formulation of the problem as a loss minimization scheme which can be solved using an end-to-end approach. The concepts of self-training and self-paced learning as applied to the proposed scheme are also presented.

2.1. Preliminaries. For a given number of sample classes, the classification task is defined as a standard softmax loss on the labeled source data as inputs x_s, y_s and the target data x_t, y_t:

$$L_c(\chi, y: \theta_c)_W = -\sum_k 1[y = k]\log P_k. \quad (1)$$

In equation (1), the aim is to produce a classifier θ_c that can correctly classify target samples at the time of testing, with minimal loss. Nonetheless, based on the assumption that there is usually a limited amount of labeled target data (potentially from only a small subset of the categories of interest), effective transfer of representations becomes limited. Consequently, a classifier abandons the less-represented class samples in the learning process, focusing only on well-represented class samples. This ultimately impedes the classifier's ability to learn robust representations. The two key issues of learning the classifier lie in an effective formulation of a score function and a robust formulation of the loss function. Again, the robustness of a learner depends on the formulation of the loss function to relieve the influence of noisy and confusing data [39]. Moreover, the works in [40, 41] proved that the optimization problem of SPL solved by the alternative optimization algorithm is equivalent to a robust loss minimization problem solved by a majorization-minimization algorithm. In view of this, the problem is formulated as minimizing the loss function:

$$\min L_c(\mathbf{W})_W = -\sum_{l=1}^{L}\sum_{n=1}^{N} \mathscr{Y}_{l,n}^{L} \log\left(P_n(W, I_l)\right) - \sum_{t=1}^{T}\sum_{n=1}^{N} \mathscr{Y}_{t,n}^{T} \log\left(P_n(W, I_t)\right). \tag{2}$$

I_l denotes the image in the source domain indexed by $l = 1, 2, 3, \ldots, L$. $\mathscr{Y}_{l,n}$ represents the true labels for the nth image ($n = 1,2, \ldots, N$) for I_l. W denotes the network weights. $P_n(w, I_l)$ is the softmax output containing the class probabilities. Similar definitions hold for $I_t, \mathscr{Y}_{t,n}$ and $p_n(w, I_t)$ during evaluation. This problem formulation is different from [35] where the number of samples is represented as union of self-labeled high-confidence samples and manually annotated samples by an active user. We further formulate to minimize the loss function in equation (3). In the case where some target labels are unavailable, these labels are assumed to be hidden and the model learns from approximate target labels $\widehat{\mathscr{Y}}$ for $\widehat{\mathscr{C}}$ (number of samples). In equation (3), $\widehat{\mathscr{Y}}$ is termed as pseudolabels:

$$\min L_c(\mathbf{W}, \widehat{\mathscr{Y}})_{W,\widehat{\mathscr{Y}}} = -\sum_{l=1}^{L}\sum_{n=1}^{N} \mathscr{Y}_{l,n}^{L} \log\left(P_n(W, I_l)\right) - \sum_{t=1}^{T}\sum_{n=1}^{N} \widehat{\mathscr{Y}}_{t,n}^{T} \log\left(P_n(W, I_t)\right). \tag{3}$$

2.2. Self-Training with Self-Paced Learning. Semisupervised learning approaches typically adopt self-training to utilize unlabeled samples [42–45]. Based on the assumption of conventional self-training, an early mistake by the learner can reinforce wrong predictions into the training set for the next training iteration. To tackle this problem, a better alternative is to resort to adding samples by adopting an "easy-to-hard" approach via self-paced learning. The principal idea in self-paced learning is generating pseudolabels from "easy" predictions on the grounds that these approximate labels are right and correctly approximate the ground truth labels, then later exploring the "hard" or less-confident pseudolabels to update the model. The self-training process used in this work is outlined in Algorithm 1. A deep CNN model is first trained with labeled samples. The model then is then used to make predictions on the unlabeled data to generate pseudolabels I_t. Similar to [30], all unlabeled samples are pseudolabeled. A novel selection algorithm with a class balancing mechanism is then used to select the nonannotated samples with the highest-confident probability predictions. These samples together with their approximated labels are added to the training set for the next training iteration. This cycle is executed iteratively until a stopping criterion is met. The overall workflow of our method is illustrated in Figure 1.

To incorporate the self-paced learning and self-training scheme, the loss function is modified as follows:

$$\begin{aligned} &\min L_c(\mathbf{W}, \widehat{\mathscr{Y}})_{W,\widehat{\mathscr{Y}}} = -\sum_{l=1}^{L}\sum_{n=1}^{N} \mathscr{Y}_{l,n}^{L} \log\left(P_n(W, I_l)\right) - \sum_{t=1}^{T}\sum_{n=1}^{N} \left[\widehat{\mathscr{Y}}_{t,n}^{T} \log\left(P_n(W, I_t)\right) + k_c \mathscr{Y}_{t,n}^{(c)}\right] \\ &\text{s.t.} \quad \mathscr{Y}_{t,n} \in \left\{e^{(i)} \in \mathbb{R}^C\right\}, k_c > 0. \end{aligned} \tag{4}$$

During training, $\mathscr{Y}$ is assigned to zero, implying that $\widehat{\mathscr{Y}}$ is ignored. To regulate the amount of pseudolabeled samples to be selected from the classes, k_c is introduced. The selection of a large quantity of pseudolabels is synonymous to a large value of k_c. Adding k_c in equation (4) introduces a class-wise bias scheme that handles the issue of class imbalance when selecting pseudolabels. The pseudolabel selection process is accomplished in two steps: (1) initialize W and minimize the loss (in equation (4)) w.r.t. $\widehat{\mathscr{Y}}_{t,n}$ and (2) set $\widehat{\mathscr{Y}}_{t,n}$ and optimize the objective function in w.r.t. W. We considered the process of executing steps 1 and 2 as a single iteration and the two steps were repeated alternatively for several iterations. The task of solving Step 1 requires a nonlinear function and as such, Step 1 was reexpressed as

$$\begin{aligned} &\min_{\widehat{\mathscr{Y}}} -\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{c=1}^{C}\left[\widehat{\mathscr{Y}}_{t,y}^{(c)} \log\left(p_n(C|wt, nI_t)\right) + k_c \widehat{\mathscr{Y}}_{t,n}^{(c)}\right] \\ &\text{s.t.} \quad \widehat{\mathscr{Y}}_{t,n} = \left[\widehat{\mathscr{Y}}_{t,n}^{(1)}, \ldots, \widehat{\mathscr{Y}}_{t,n}^{(c)}\right] \in \left\{e^{(i)} \in \mathbb{R}^C\right\}, k_c > 0. \end{aligned} \tag{5}$$

The introduction of a class-wise bias by normalizing class-wise confidence scores distinguishes this formulation from the one proposed in [21] where the authors adopted an L_1 regularizer in a bid to avoid the scenario where most of

input: Deep Learning Network $D(w)$, unlabeled Images I_t, amount K_c
output: Trained Classifier (C)
Train a deep network $D(w)$ with labeled samples I_l
for $k \longleftarrow 1$ **to** N **do**
  Test and predict on unlabeled samples I_t;
  Generate pseudolabels for I_t using predictions;
  Select K_c pseudolabeled samples after filtering out balancing class-wise scores
  Augment labeled training set $(I_l + K_c(I_t))$ with selected K_c pseudolabeled samples
  Retrain $D(w)$ with I_l and K_c pseudolabeled samples $(I_l + K_c(I_t))$
end
$C = updated(D(w))$;
Return C

Algorithm 1: Self-paced learning workflow.

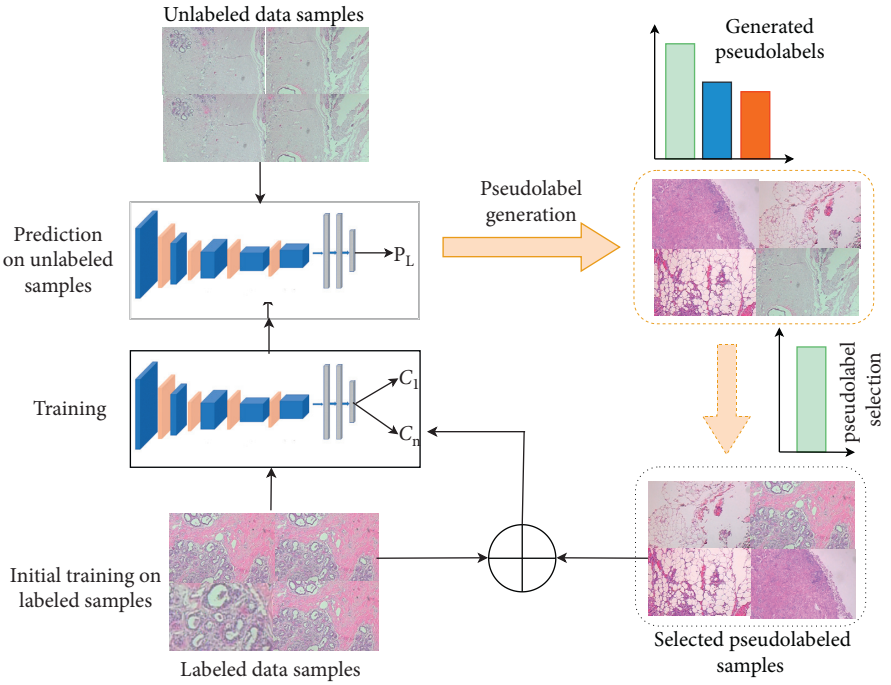


Figure 1: Workflow of the proposed approach. A deep CNN model is first trained with labeled data samples. The trained model is then evaluated on unlabeled data to generate pseudolabels for the unlabeled data. A pseudolabel selection algorithm that integrates a class balancing mechanism is used to select pseudosamples that have the highest confidence probability confidence score. The selected samples together with their pseudolabels are used to augment the training sample for the next training iteration and the cycle is repeated iteratively until a stopping criterion is met.

the pseudolabels are ignored. In solving the pseudolabel framework optimizer, the work in [21] utilized the solver expressed in the following equation:

$$\widehat{\mathcal{Y}}_{t,y}^{(c*)} = \begin{cases} 1, & \text{if } c = \arg\max p_n(c|w, I_t),\ p_n(c|w, I_t) > \exp(-k) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

With such a formulation, the process of generating and selecting pseudolabels hinges on the output probability $(p_n(c|wt, nI_t))$. Inherently, such an approach does not handle the issue of class imbalance. To resolve this, equation (3) is reexpressed as follows:

$$\begin{aligned} \min L_c(W, \widehat{\mathcal{Y}})_{W,\widehat{\mathcal{Y}}} &= -\sum_{l=1}^{L}\sum_{n=1}^{N} Y_{l,n}^{L} \log\left(P_n(W, I_l)\right) \\ &\quad - \sum_{u=1}^{T}\sum_{n=1}^{N}\sum_{c=1}^{C}\left[\widehat{\mathcal{Y}}_{t,n}^{T} \log\left(P_n(W, I_t)\right) + k_c \widehat{\mathcal{Y}}_{t,n}^{(c)}\right] \\ \text{s.t.}\, \widehat{\mathcal{Y}}_{t,n} &= \left[\widehat{\mathcal{Y}}_{t,n}^{(1)}, \ldots, \widehat{\mathcal{Y}}_{t,n}^{(c)}\right] \in \left\{e^{(i)} \in \mathbb{R}^C\right\}, k_c > 0. \end{aligned} \tag{7}$$

Minimizing the optimization framework in equation (7) was accomplished by using the loss function in equation (5) but with a solver that incorporates the class-wise normalizing term (different from the one proposed in [21]) expressed as

$$\widehat{Y}_{u,y}^{(c*)} = \begin{cases} 1, & \text{if } c = \arg\max \dfrac{p_n(c|w, I_u)}{\exp(-k_c)}, \dfrac{p_n(c|w, I_u)}{\exp(-k_c)} > 1 \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The process of generating and selecting pseudolabeled samples is dependent on the normalized class-wise output $(p_n(c|w, I_u))/(\exp(-k_c))$ in equation (8). Using the normalized output ensures a balance towards classes with relatively low score but with a high intraclass confidence score during the process of assigning pseudolabels to an unlabeled sample.

To regulate the amount of pseudolabeled samples to be selected to update the model in each training iteration, K_c is set using the process in Algorithm 2. In finding and fixing a value for K_c, the algorithm ranks the class C probabilities on all the image samples predicted as class C. K_c is set such that $\exp(-K_c)$ is equivalent to the probability ranked at iteration $(p * N_c)$, with N_c being the number of images predicted as class C. For each unlabeled sample, the maximum output probability M was taken in descending order and these probabilities are sorted out across all samples. Optimizing the pseudolabels resulted in the $p \times 100\%$ most confident pseudolabeled samples to be used in training the model (where p is a scaled proportion between [0, 1]). Such a scheme ensures that the probability ranked at $p \times 100\%$ is taken independently from each class to (1) threshold the confidence scores and (2) normalize the confidence scores. p is first initialized with 10% of the most confident predictions and at each additional round, the top 5% is added to the next pseudolabel generation and selection process.

3. Materials and Experiments

3.1. Dataset. We have carried out experiments on the BreakHis dataset [18]. The BreakHis dataset contains microscopic biopsy images of benign and malignant breast tumors totaling 7909 images. The image samples were generated from breast tissue biopsy slides, stained with hematoxylin and eosin (HE). Each image has a pixel size of 700×460 (in PNG format), with a 3-channel RGB, and 8-bit depth in each channel. The benign and malignant classes are each further subdivided into four distinct types. The subtypes for the benign class are adenosis, fibroadenoma, phyllodes tumors, and tabular adenoma. The malignant class subtypes are ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. The images are obtained using four magnification factors −40X, 100X, 200X, and 400X. The images exhibit fine-grained characteristics with only subtle differences between images from different classes as well as high coherency, which is typical of cancerous cells. These factors, compounded with the fact that images in the same class have different contrasts and resolutions, make the BreakHis dataset challenging, not to mention the high imbalance in subtype classes (2,480 images belong to the benign class and 5,429 images belong to the malignant class). Figure 2 shows sample images from each subtype class and Table 1 shows the distribution of images per each class.

3.2. Experimental Settings. The pretrained Inception_ResNetV2 [46], a variant of the Inception_V3 model [47], was used as the baseline model for all experiments. Inception_ResNetV2 is able to greatly improve classification and recognition performance at low computational costs. Input images are resized to 299×299 before being fed to the model. At the fully supervised learning phase, the baseline model is fine-tuned to initialize the model weights and also reduces variance. Fine-tuning of pretrained models has demonstrated to be an effective approach for achieving significantly higher results even on small-scale data. For the supervised learning phase, the model is trained for a total of fifty (50) epochs using the Adam optimizer [48], $\beta_1 = 0.9$, $\beta_2 = 0.99$ and an initial learning rate of 0.001 which is decayed via a polynomial decay scheduling (expressed in equation (9)). A polynomial decay scheduling allows the learning rate to decay over a fixed number of epochs:

$$\alpha = \text{initLR} * \left(1 - \frac{\text{epoch}}{T_{\text{epochs}}}\right)^{p}, \tag{9}$$

initLR is the base learning rate, T_{epochs} is the total number of epochs, and p is the exponential power, which is set to 1. The model is trained with a batch size of 32. Random rotation with a range of 90° and horizontal flipping have been implemented as data augmentation techniques to help combat overfitting. For the self-training phase, the model is also retrained with hyperparameters for top K_c using 5%, 10%, and 20% of the pseudolabeled samples of the unlabeled data. 70% of the data is used as training data and 30% is added to the test samples to be used as the unlabeled data for the self-training scheme. The training data was further split into 70 : 30 percent ratio as training and validation data, respectively. The model is trained for a total of 5 iterations during the semisupervised phase. We experimented with 5, 8, and 10 iterations and realized that not only did the 8 and 10 iterations take too much time to train, they also did not contribute significantly to the accuracy of the model compared to training for 5 iterations. To efficiently optimize training time, we decided to train for 5 iterations as this resulted in excellent accuracy within a limited time. Each experiment is repeated three times and the results are averaged. The iterations were stopped when there was no further improvement in accuracy.

The proposed approach does not add extra computational overhead during training, allowing training to be completed in an efficient manner. The averaged total training time for all experiments is shown in Tables 2 and 3, respectively. All experiments are carried out using Keras (version 2.2.4) with TensorFlow backend (version 1.12) and CUDA 9.0. Two RTX 2080 graphic cards, each with 8 GB

input: Deep CNN $D(w)$, unlabeled samples I_t, selected pseudolabels p
output: K_c
for $t \longleftarrow 1$ to T **do**
  $P_{I_t} = D(w, I_t)$;
  $LP_{I_t} = argmax(P, axis = 0)$;
  $MP_{I_t} = \max(P, axis = 0)$;
  $M = [M, from - matrix - to - vector(MP_{I_t})]$**for** $c \longleftarrow 1$ **to** C **do**
    $MP_c, I_t = MP_{I_t}(LP_{I_t} == c)$;
    $M_c = [M_c, Matrix - to - vector(MP_c, I_t)]$
  end
end
  for $c \longleftarrow 1$ **to** C **do**
  $M_c = sort(M_c, order = descending)$;
  $len_{c,th} = length(M_c) \times p$;
  $K_c = -\log(M_c[len_{c,th}])$
end
$return(K_c)$

Algorithm 2: Determining K_c.

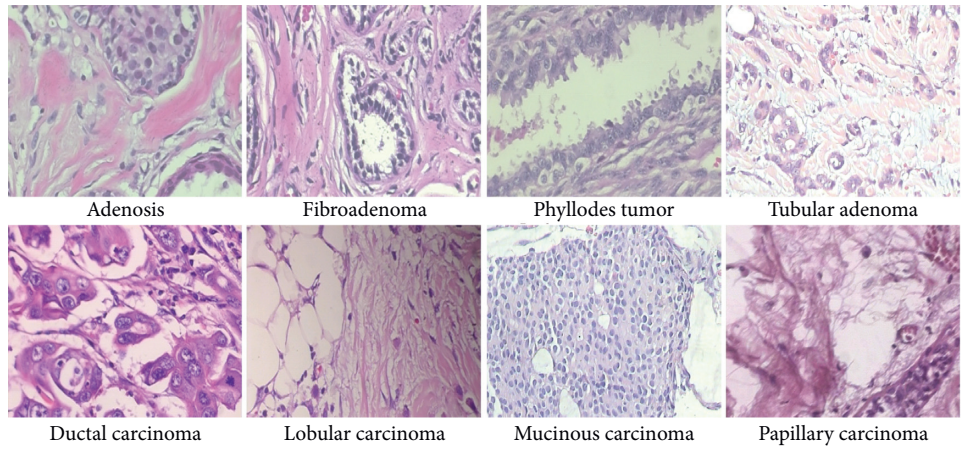


Figure 2: Sample image from each of the eight cancer subtypes in the BreakHis dataset. The images have subtle differences across classes due to their fine-grained nature, with different contrast and resolutions. These characteristics, coupled with the high coherency of the cancerous cells, make the dataset a challenging one. The images are obtained at a magnification factor of 200X.

Table 1: The distribution of images per individual subtype classes of the BreakHis histopathological images dataset.

Class	Subtype	Magnification factors			
		40X	100X	200X	400X
Benign	Adenosis	114	113	111	106
	Fibroadenoma	193	260	264	137
	Phyllodes tumors	149	150	140	130
	Tabular adenoma	109	121	108	115
Malignant	Ductal carcinoma	864	903	896	788
	Lobular carcinoma	156	170	163	137
	Mucinous carcinoma	205	222	196	169
	Papillary carcinoma	145	142	135	138

The distribution shows unequal number of image distribution per classes, resulting in class imbalance which makes the dataset a challenging one.

Table 2: Average training times for the binary classification task based on the amount of selected pseudolabels.

% of pseudolabels	40X	100X	200X	400X
K(top-5) pseudolabels	1 hour 59 min	2 hours 2 min	2 hours 4 min	1 hour 50 min
K(top-10) pseudolabels	1hour 56 min	2 hours 2 min	1 hour 59 min	1 hour 50 min
K(top-20) pseudolabels	1 hour 59 min	2 hours 3 min	2 hours	1 hour 53 min
All pseudolabels	1 hour 58 min	2 hours 4 min	1 hour 57 min	1 hour 49 min

min represents minutes.

memory and a 32 GB RAM, served as the hardware platforms. The evaluation metrics used in accessing the model were classification accuracy, precision, recall, F1-score, and confusion matrix. These parameters are related to the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates, respectively. True positive measures how correctly a classifier predicts the positive class. True negative measures how correctly a classifier predicts the negative class. False positive measures how, incorrectly, a

Table 3: Average training times for the multiclass classification task based on the amount of selected pseudolabels.

% of pseudolabels	40X	100X	200X	400X
K(top-5) pseudolabels	2 hours	2 hours 5 min	2 hours 1 min	1 hour 47 min
K(top-10) pseudolabels	1hour 58 min	2 hours 5 min	1 hour 57 min	1 hour 49 min
K(top-20) pseudolabels	2 hours 1 min	2 hours	2 hours 2 min	1 hour 49 min
All pseudolabels	2 hours	2 hours 5 min	2 hours	1 hour 49 min

min represents minutes.

classifier predicts the positive class. False negative measures how, incorrectly, a classifier predicts the negative class.

4. Results and Discussion

The proposed scheme was evaluated using the top 5%, 10%, and 20% pseudolabeled samples. For purposes of reporting and investigation, we also report on values obtained when all pseudolabeled samples (100%) were used. We present and discuss results for both binary and multiclass classification tasks.

4.1. Binary Classification. The experimental outcomes for the binary classification task are shown in Table 4. For images with magnification factor of 40X, the best accuracy result was 99.52% when the top-10% pseudolabeled samples were selected. Similarly, for a magnification factor of 100X, the best accuracy result was 99.44% with the top-5% pseudolabeled samples. Using the top-10% pseudolabeled samples resulted in 99.48% accuracy for images with a magnification factor of 200X, and using the top-10% yielded an accuracy result of 99.47% with images scanned at 400X.

The generation and selection of the top K_c pseudolabeled samples via the proposed schemed was a vital key in controlling and determining the amount of pseudolabeled samples to be selected in updating the model at the next iteration. The selection scheme, coupled with the self-paced learning and self-training approach ensured that classes with the least representations which would have otherwise been ignored, was still selected and added to the training samples. This proved to be an effective and efficient step in the learning process. Again, the results in Table 4 show that selecting the top K_c pseudolabels proved to be a more effective approach rather than using all the pseudolabeled samples. The accuracy results obtained with the proposed approach show significant accuracy gains.

The accuracy and loss plots for 40X and 100X are shown in Figures 3 and 4 denotes plots for 200X and 400X, respectively. When training deep networks, overfitting remains a vital issue that needs to be addressed as it affects the ability of a trained model to generalize well on new data. It is observed from the plots that both accuracy and loss values were unstable until after epoch thirty (during the supervised learning stage). Values kept bouncing within different intervals from the start of training till the epoch thirty. We attribute this to the distance disparity between the source and target data. In fine-tuning a pretrained model on a secondary task, there is the assumption that the source and target domains are related to each other. However, in cases where this assumption is not met, brute-force transfer learning may not be successful and even in the worst case, degrading learning performance in the target domain [49].

The pretrained model used as the baseline model was trained on the ImageNet dataset (which consists of natural images) as against the BreakHis dataset which contains breast cancer histopathological images. As such, at the start of supervised training stage, the model begins to learn the relatively new patterns from the target domain (breast cancer images) resulting in the spikes as depicted in the plots. However, past epoch thirty, a drastic drop in loss value is observed and the accuracy values increase steadily. At the end of epoch fifty, the loss value is greatly reduced and the training and validation accuracy (for both the supervised learning stage and the self-training stage) are almost aligned. This is an indication that the proposed approach also effectively curbs overfitting. The imbalanced nature of the BreakHis dataset implies that accuracy alone cannot be used to access the performance of the model. Results for precision, recall, and F1-score values are also presented in Table 5. The confusion matrices are also presented in Figure 5. The BreakHis dataset contains more samples for the malignant class compared to the benign class, and this is also reflected in the confusion matrices. Nonetheless, the selection process together with the class balancing framework adopted in this work ensured the fact that the model accurately classified the respective classes with minimal misrepresentations.

4.2. Multiclass Classification. The accuracy results for the multiclass classification are summarized in Table 6. For images scanned at 40X, the highest accuracy obtained was 94.28% when the top-10% pseudolabels were selected. For 100X, the best accuracy was 93.84% when the top-20% pseudolabels were selected. Selecting the top-5% pseudolabels yielded an accuracy of 94.93% for images scanned at a magnification factor of 200X. For images scanned at a magnification factor of 400X, the best accuracy was 93.75% when the top-10% pseudolabels were selected. Similar to the binary classification task, selecting the top K_c pseudolabels to augment the training samples in the next training iteration proved to be more effective than selecting all the pseudolabels. This outcome further rubber-stamps the significance of K_c in the proposed approach.

The plots for loss and accuracy (for images scanned at 40X and 100X) are shown in Figure 6 and the corresponding plots for 200X and 400X are shown in Figure 7. The nature of the plots follow from the explanations provided for the

Table 4: Accuracy (%) performance for binary classification. Baseline indicates that the model was fine-tuned with labeled samples only. K_c (top-N) indicates the portion of the most confident pseudolabels used. Best results are indicated in italics.

ST approach	40X	100X	200X	400X
Baseline	97.14 ± 0.33	98.22 ± 0.40	98.55 ± 0.57	98.43 ± 0.44
K_c(Top-5%) pseudolabels	99.28 ± 0.6	*99.44 ± 0.41*	99.03 ± 0.34	99.04 ± 0.73
K_c(Top-10%) pseudolabels	*99.52 ± 0.33*	98.85 ± 0.32	*99.48 ± 0.30*	*99.47 ± 0.37*
K_c(Top-20%) pseudolabels	99.27 ± 0.37	97.95 ± 0.01	98.79 ± 0.68	98.92 ± 0.22
All pseudolabels	98.09 ± 0.21	98.20 ± 0.13	98.5 ± 0.72	98.69 ± 0.58

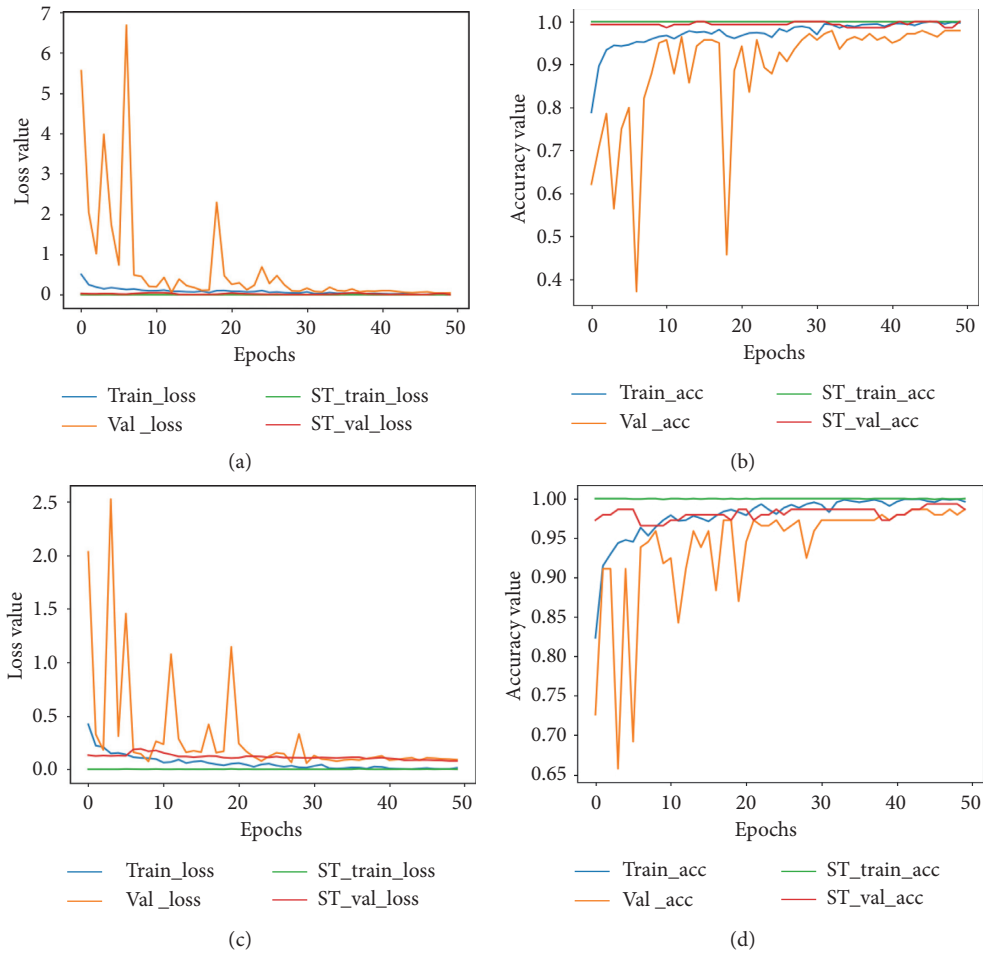


Figure 3: Accuracy plot for images scanned at 40X and 100X for the binary classification task. (a) The loss plot for 40X and (b) the corresponding accuracy plot. (c) The loss plot for 100X and (d) the corresponding accuracy plot. ST represents the self-training plot.

binary classification plot. The precision, recall, and F1-score values are provided in Table 7 and the confusion matrices for all magnification factors are provided in Figure 8.

The confusion matrices also bring out the imbalance in the dataset. The ductal carcinoma class has more samples than the remaining classes with the adenosis class having the least number of samples. As a result, these two classes represent the most and least number of samples, as depicted in Figure 8. Again, the subtle nature of the appearance of the different images per different classes also does pose challenges for models in accurately discriminating between classes. In [23], the authors pointed out this difficulty, especially when discriminating between ductal carcinoma and lobular carcinoma as well as fibroadenoma and tabular adenoma. However, from the confusion matrices, it is observed that such misrepresentations are effectively handled by the proposed approach. Between ductal carcinoma and lobular carcinoma, an average of four samples are misrepresented while between fibroadenoma and tubular adenoma, only two samples are misrepresented for images scanned at a magnification factor of 200X.

The accuracy, precision, recall, and F1-score values as well as the confusion matrices all show the effectiveness of using K_c in determining the proportions of pseudolabels to be used in updating the model in each training iteration and also prove that adding samples in an "easy-to-hard"

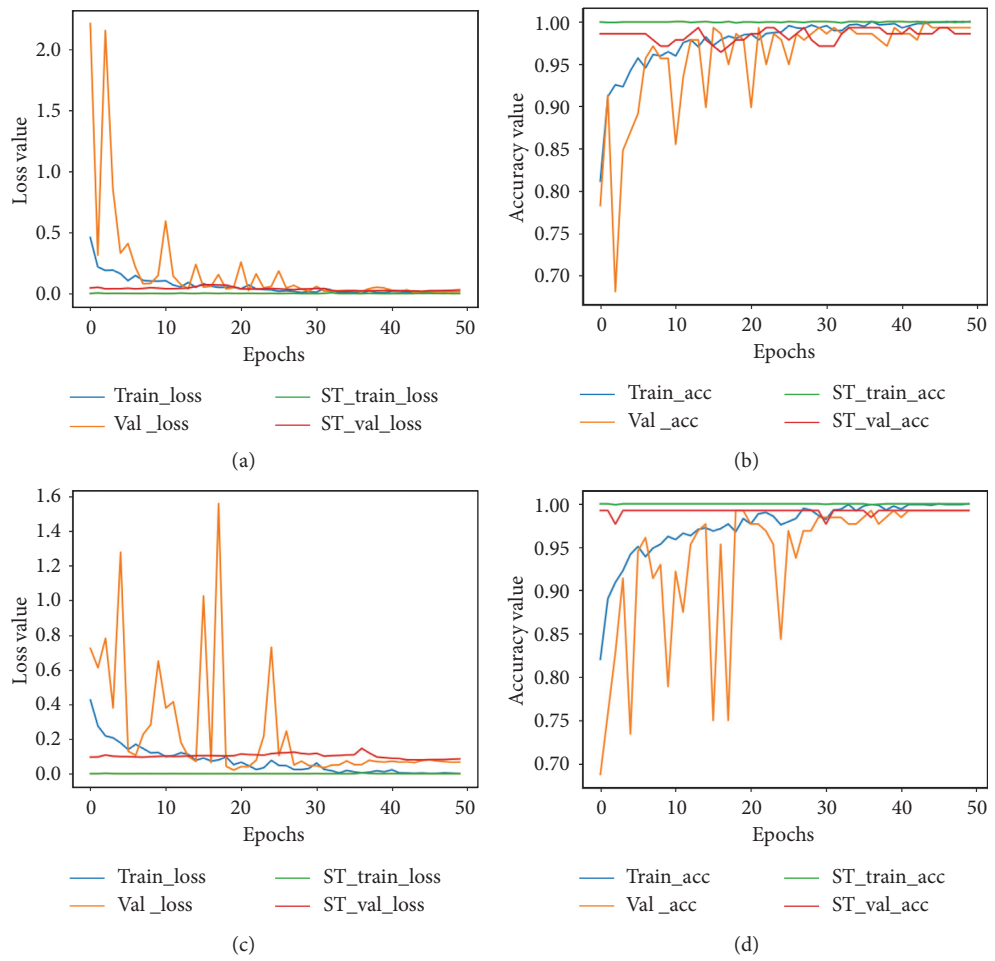


Figure 4: Accuracy plot for images scanned at 200X and 400X for the binary classification task. (a) The loss plot for 200X and (b) the corresponding accuracy plot. (c) The loss plot for 400X and (d) the corresponding accuracy plot. ST represents the self-training plot.

Table 5: Precision (Prec.), recall (R), and F1-score (F1) values for binary classification.

Mag. factor	% of pseudolabels	Prec. (%)	R (%)	F1 (%)
40X	K_c(top-5%)	99.50	99.23	99.38
	K_c(top-10%)	99.89	99.79	99.81
	K_c(top-20%)	99.50	99.21	99.36
	All pseudolabels	98.72	98.63	98.49
100X	K_c(top-5%)	99.73	99.58	99.69
	K_c(top-10%)	99.28	99.17	99.23
	K_c(top-20%)	98.62	98.24	98.71
	All pseudolabels	99.12	99.06	99.19
200X	K_c(top-5%)	99.43	98.91	99.18
	K_c(top-10%)	99.84	99.80	99.49
	K_c(top-20%)	99.27	99.00	99.13
	All pseudolabels	99.18	99.10	99.22
400X	K_c(top-5%)	99.40	99.17	99.20
	K_c(top-10%)	99.85	99.77	99.54
	K_c(top-20%)	99.25	99.18	99.21
	All pseudolabels	99.20	99.00	99.14

approach ensures that even the least-represented samples are still considered in the training process. Overall, these schemes resulted in the model being very versatile and robust even in the face of the similarities and coherence between the images samples in the dataset.

4.3. Comparison with Other Works. We compare the performance of the proposed approach with other works mentioned in the literature as shown in Table 8 for the binary classification task) and Table 9 (for the multiclass classification task), respectively. All these underlisted state-of-the-art methods were evaluated on the BreakHis dataset, offering a fair comparison and assessment with the proposed approach in this work. The work in [23] used a CNN model consisting of five convolutional layers and two fully connected layers for both binary and multiclass classification tasks. Using an ensemble method, the authors report accuracy of 98.33%, 97.12%, 97.85%, and 96.15% for magnification factors 40X, 100X, 200X, and 400X for the binary classification task. For the multiclass classification, they reported accuracy of 88.23%, 84.64%, 83.31%, and 83.39% for magnification factors of 40X, 100X, 200X, and 400X.

In [24], the authors proposed a structured deep learning model for classifying breast cancer histopathological images. In their work, the authors considered the feature space similarities of histopathological images by leveraging intra-

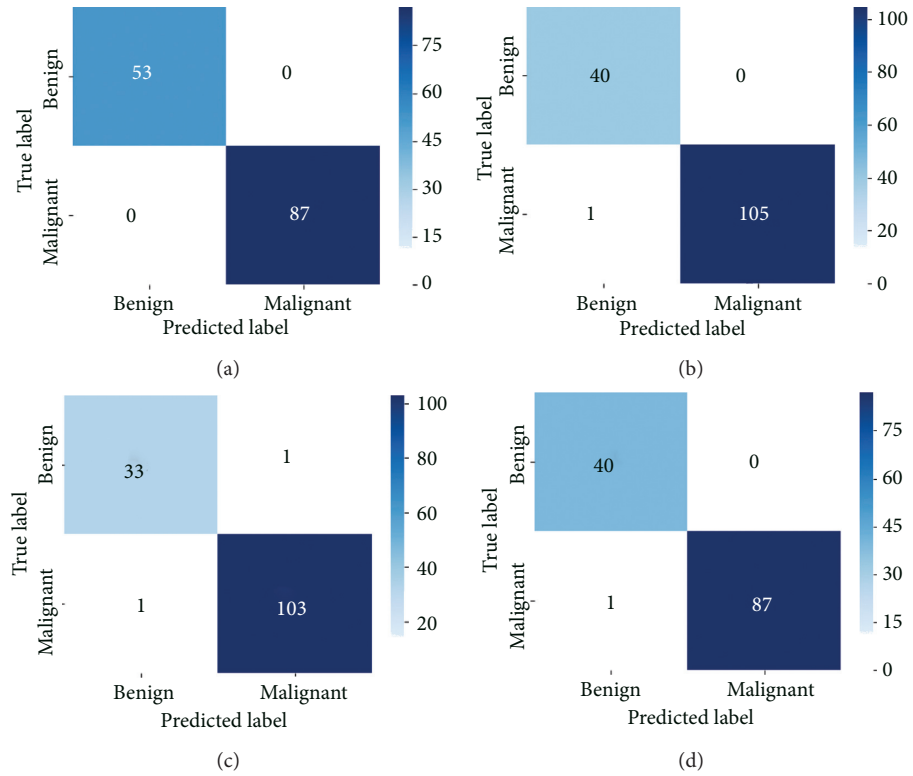


Figure 5: Confusion matrix for binary classification. (a) 40X. (b) 100X. (c) 200X. (d) 400X.

Table 6: Accuracy (%) for multiclass classification.

ST approach	40X	100X	200X	400X
Baseline	91.42 ± 0.54	89.04 ± 0.70	90.07 ± 0.21	90.70 ± 0.63
K_c(top-5%) pseudolabels	94.27 ± 0.28	91.78 ± 0.61	*94.93 ± 0.17*	92.97 ± 0.37
K_c(top-10%) pseudolabels	*94.28 ± 0.29*	92.46 ± 0.48	94.32 ± 0.22	*93.75 ± 0.72*
K_c(top-20%) pseudolabels	94.14 ± 0.14	*93.84 ± 0.35*	91.48 ± 0.28	92.19 ± 0.16
All pseudolabels	92.87 ± 0.71	90.41 ± 0.63	92.19 ± 0.38	91.40 ± 0.11

Baseline indicates that the model was fine-tuned with labeled samples only. K_c (top-N) indicates the portion of the most confident pseudolabels used. Best results are indicated in italics.

and interclass labels as prior knowledge. They also adopted a data augmentation scheme that generated more data for the model during training. Using a pretrained deep CNN model as their base network, the authors reported accuracy of 95.8%, 96.9%, 96.7%, and 94.9% for the binary classification task. For the multiclass task, they reported accuracy of 92.8%, 93.9%, 93.7%, and 92.9% for magnification factors of 40X, 100X, 200X, and 400X, respectively. It can be observed that their approach yielded a 0.06% gain in accuracy for images scanned at 100X for the multiclass task compared to our approach. The data augmentation approach used in their work amassed more data for model during the fine-tuning stage compared to our approach and their overall approach was a supervised one (meaning only labeled data was used) as opposed the semisupervised fashion in ours (SSL dwells on the assumption that there are more unlabeled samples than labeled samples [27]). That notwithstanding, our approach yielded significant accuracy improvements for all the other magnification factors.

In [51], the authors proposed a novel L-Isomap-aided manifold learning and stacked sparse autoencoder framework for a robust BC classification using HIs. The authors reported accuracy of 96.8%, 98.1%, 98.2%, and 97.5% for images with magnification factors 40X, 100X, 200X, and 400X, respectively. In [50], the authors used a CNN model to extract local and frequency domain information from input images for classifying breast cancer images on the BreakHis dataset. They report accuracy of 94.40%, 95.93%, 97.19%, and 96.00% for the binary classification task. These algorithms mentioned in the literature only utilize supervised learning approaches.

In this work, we have used 70% of the data for training at the supervised learning stage and the remaining 30% was added to the test set which was used as unlabeled data for the self-training stage. The selection of the most confident pseudolabeled samples to augment the training sample has been proven effective in providing the model with reliable

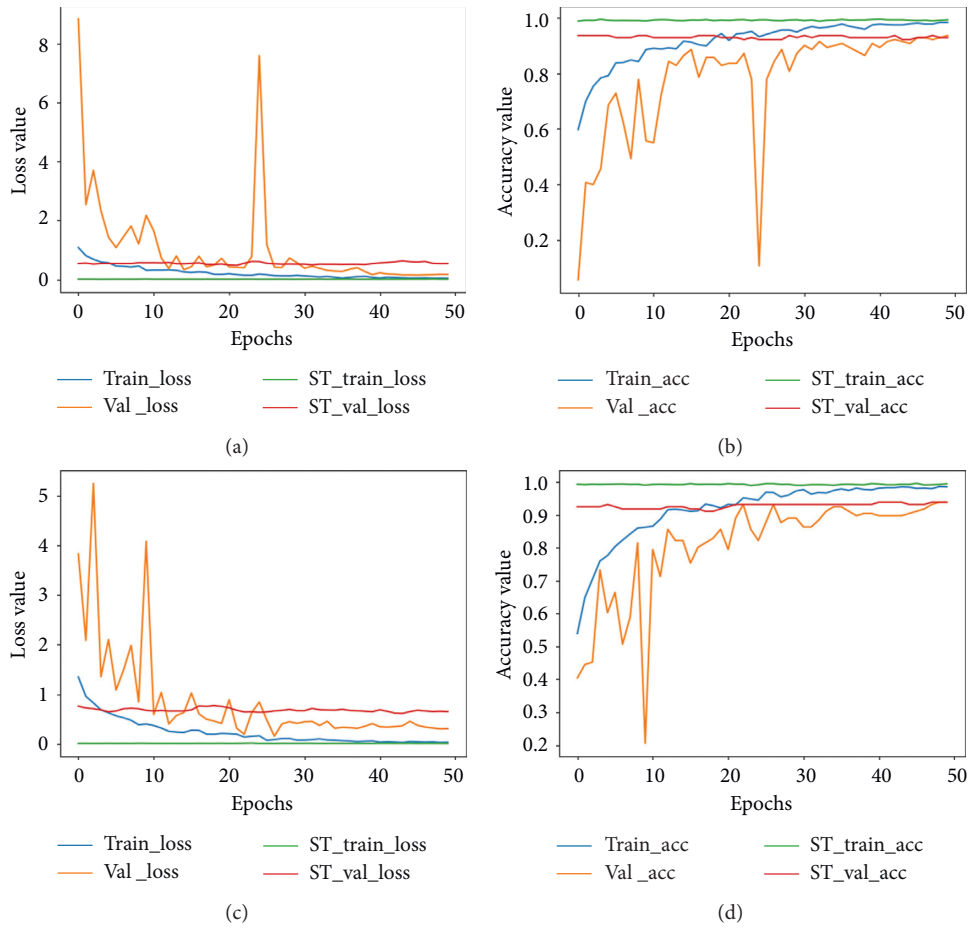


Figure 6: Accuracy and loss plot for images scanned at 40X and 100X for the multiclass classification. (a) The loss plot for 40X and (b) the corresponding accuracy plot. (c) The loss plot for 100X and (d) is the corresponding accuracy plot. ST represents the self-training plot.

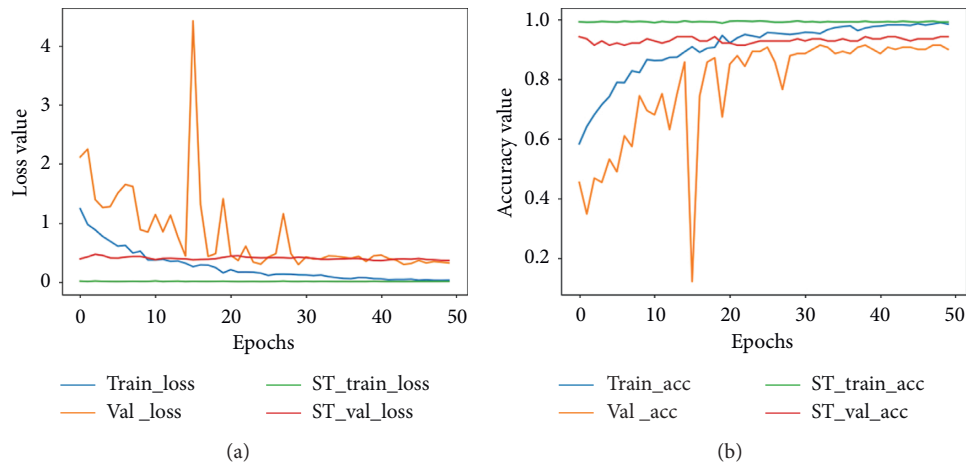
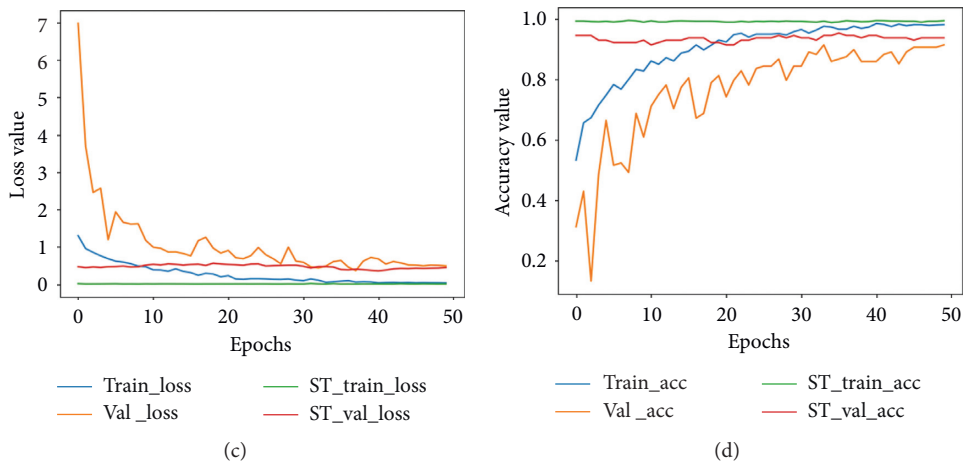


Figure 7: Continued.

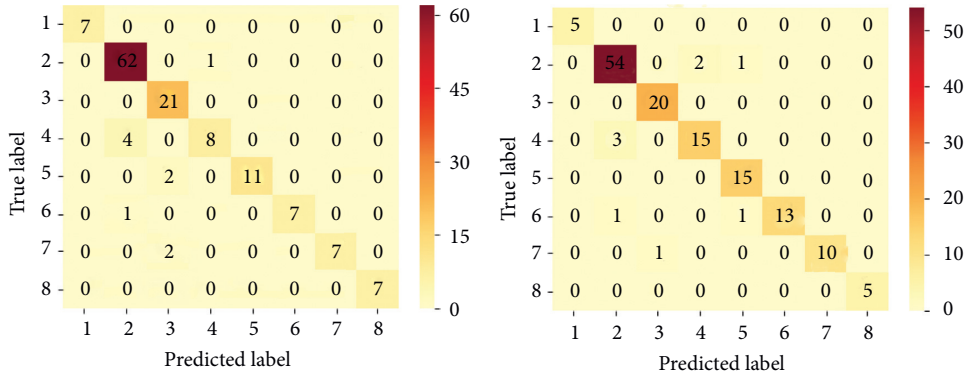


(c) (d)

Figure 7: Accuracy plot for images scanned at 200X and 400X for the multiclass classification task. (a) The loss plot for 200X and (b) the corresponding accuracy plot. (c) The loss plot for 400X and (d) the corresponding accuracy plot. ST represents the self-training plot.

Table 7: Precision (Prec.), recall (R), and F1-score (F1) values for multiclass classification.

Mag. factor	% of Pseudolabels	Prec. (%)	R (%)	F1 (%)
40X	K_c(top-5%)	94.96	94.71	94.67
	K_c(top-10%)	95.15	94.78	94.80
	K_c(top-20%)	94.63	94.55	94.59
	All pseudolabels	93.25	93.0	93.21
100X	K_c(top-5%)	91.85	91.71	91.89
	K_c(top-10%)	93.14	92.85	92.38
	K_c(top-20%)	94.24	94.71	94.33
	All pseudolabels	90.63	90.39	90.51
200X	K_c(top-5%)	95.85	95.90	95.56
	K_c(top-10%)	95.47	94.91	95.32
	K_c(top-20%)	91.44	91.07	91.64
	All pseudolabels	92.65	92.0	92.75
400X	K_c(top-5%)	93.48	93.23	93.51
	K_c(top-10%)	94.36	94.28	94.38
	K_c(top-20%)	92.69	92.14	92.32
	All pseudolabels	90.63	90.57	90.41



(a) (b)

Figure 8: Continued.

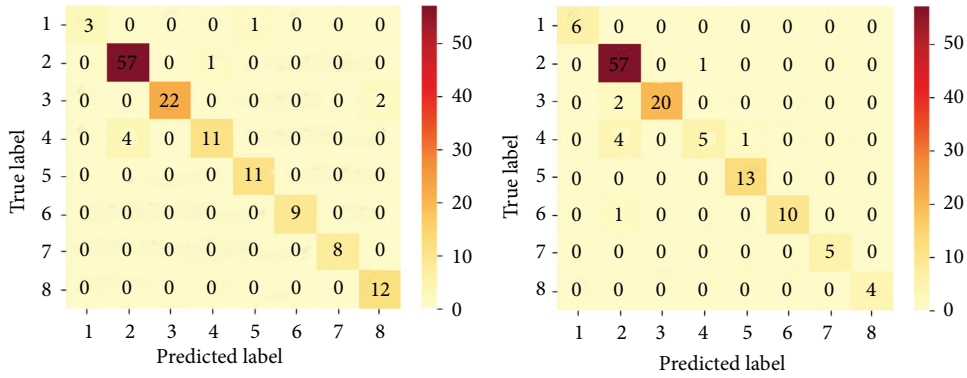


Figure 8: Confusion matrix for multiclass classification for the respective magnification factors. The imbalance in the sample distribution is evident in the plot. Nonetheless, there are not so many misrepresentations among classes. Order of class names: 1: adenosis, 2: ductal carcinoma, 3: fibroadenoma, 4: lobular carcinoma, 5: mucinous carcinoma, 6: papillary carcinoma, 7: phyllodes tumor, and 8: tubular adenoma. (a) 40X. (b) 100X. (c) 200X. (d) 400X.

Table 8: Accuracy comparison with some state-of-the-art algorithms for the binary classification task on the BreakHis dataset.

Ref.	Mag. fac.	Acc. (%)	Prec. (%)	R (%)	F1 (%)
Nahid and Kong [50]	40X	94.40	94.00	96.00	95.00
	100X	95.93	98.00	96.36	97.00
	200X	97.19	98.00	98.20	98.00
	400X	96.00	95.00	97.79	96.00
Han et al. [24]	40X	95.8 ± 3.1	—	—	—
	100X	96.9 ± 1.9	—	—	—
	200X	96.7 ± 2.0	—	—	—
	400X	94.9 ± 2.8	—	—	—
Pratiher and Chattoraj [51]	40X	96.8	—	—	—
	100X	98.1	—	—	—
	200X	98.2	—	—	—
	400X	97.5	—	—	—
Bardou et al. [23]	40X	98.33	97.80	97.57	97.68
	100X	97.12	95.58	96.98	97.77
	200X	97.85	95.61	99.28	97.41
	400X	96.15	97.54	96.49	97.07
K_c(top-10% pseudolabels)	40X	99.52 ± 0.33	99.50	99.23	99.38
K_c(top-5% pseudolabels)	100X	99.44 ± 0.41	99.73	99.58	99.69
K_c(top-10% pseudolabels)	200X	99.48 ± 0.30	99.84	99.80	99.49
K_c(top-10% pseudolabels)	400X	99.47 ± 0.37	99.85	99.77	99.54

Acc. denotes the accuracy, Prec. is the precision, R is the recall, and F1 is the F1-score.

Table 9: Accuracy comparison with some state-of-the-art algorithms for the multiclass classification task on the BreakHis dataset.

Ref.	Mag. fac.	Acc. (%)	Prec. (%)	R (%)	F1 (%)
Han et al. [24]	40X	92.8 ± 2.1	—	—	92.9
	100X	93.9 ± 1.9	—	—	88.9
	200X	93.7 ± 2.2	—	—	88.7
	400X	92.9 ± 1.8	—	—	85.9
Bardou et al. [23]	40X	88.23	84.27	83.79	83.74
	100X	84.64	84.29	84.48	84.31
	200X	83.31	81.85	80.83	80.48
	400X	83.98	80.84	81.03	80.63
K_c(top-10% pseudolabels)	40X	94.28 ± 0.29	95.15	94.78	94.80
K_c(top-20% pseudolabels)	100X	93.84 ± 0.41	94.24	94.71	94.33
K_c(top-5% pseudolabels)	200X	94.93 ± 0.17	95.85	95.90	95.56
K_c(top-10% pseudolabels)	400X	93.75 ± 0.72	94.36	92.28	94.38

Acc. denotes the accuracy, Prec. is the precision, R is the recall, and F1 is the F1-score.

samples, and ultimately expanding the training set, thereby making more data available to the model (to satisfy the hunger of deep models for more data). The effectiveness of the proposed method is evident in the results obtained, which depict significant accuracy improvements compared to the abovementioned methods which are mostly supervised learning approach where only labeled data was used. The proposed algorithm has been tested on breast cancer histopathological images since it is in line with our research objective. Therefore, we are quick to add that, the significance of the proposed algorithm is not limited or specifically designed for breast cancer classification. Based on the results obtained, we are confident that this algorithm can be extended to other classification tasks in medical imaging or computer vision that seek to employ semisupervised learning techniques in solving various tasks.

5. Conclusion

Obtaining a significant amount of well-labeled data in the medical domain is a challenging task and more tedious is the task of accurately providing labels to data. In this work, we have proposed a semisupervised learning scheme that integrates self-paced learning paradigm and self-training for training a model on both labeled and unlabeled data. Self-paced learning plays a vital role in curbing the issue of mistake reinforcement, where wrongly generated pseudo-labels are reinforced into the training sample. In the light of selecting pseudolabels with the most confident probabilities, we show a novel selection algorithm was proposed to present the CNN model with only the most confident pseudolabels. Experimental results obtained using the top 5%, 10%, and 20% generated pseudolabels for training showed significant accuracy improvements for both binary and multiclass classification task when compared with state-of-the-art approaches. For future work, we intend to incorporate diversity into the self-paced learning scheme and as well as incorporate the similarities in feature space of histopathological images. A combination of these elements into the self-paced learning scheme will result in a versatile and robust learner.

Data Availability

The data used in this work are available from [18] (DOI: https://doi.org/10.1109/TBME.2015.2496264).

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] American Cancer Society, *Breast Cancer Facts and Figures*, American Cancer Society, Atlanta, GA, USA, 2019.

[3] S. Z. Ramadan, "Methods used in computer-aided diagnosis for breast cancer detection using mammograms: a review journal of healthcare engineering," *Journal of Healthcare Engineering*, vol. 2020, Article ID 9162464, 21 pages, 2020.

[4] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 236–251, 2009.

[5] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 3, pp. 538–556, 2012.

[6] G. Mariscotti, N. Houssami, M. Durando et al., "Digital breast tomosynthesis (DBT) to characterize MRI-detected additional lesions unidentified at targeted ultrasound in newly diagnosed breast cancer patients," *European Radiology*, vol. 25, no. 9, pp. 2673–2681, 2015.

[7] J. G. Elmore, G. M. Longton, P. A. Carney et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *The Journal of the American Medical Association*, vol. 313, no. 11, pp. 1122–1132, 2015.

[8] T. Ahmad, G. Aresta, E. Castro et al., "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, Article ID e0177544, 2017.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 23–28, Columbus, OH, USA, June 2014.

[10] K. Malik, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[11] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the Conference on Computer Vision - ECCV 2018*, p. 297, Munich, Germany, October 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations*, pp. 770–778, San Diego, CA, USA, May 2015.

[14] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[15] B. J. Suk, P. Korfiatis, Z. Kline, and T. L. Kline, "Machine learning for medical imaging," *RadioGraphics*, vol. 37, no. 2, pp. 505–515, 2017.

[16] Y. Akkus, B. Zhang, F. Coenen, J. Lu, and W. Lu, "One-class kernel subspace ensemble for medical image classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 13, 2014.

[17] Y. Xiao, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles," *Machine Vision and Applications*, vol. 24, no. 7, pp. 1405–1420, 2013.

[18] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[19] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] K. Hinton, S. E. A. Raza, Y.-W. Tsang et al., "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[21] S. K. Asare, F. You, and O. T. Nartey, "A robust pneumonia classification approach based on self-paced learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 83–89, 2020.

[22] O. T. Tettey, G. Yang, S. K. Asare, J. Wu, and L. N. Frempong, "Robust semi-supervised traffic sign recognition via self-training and weakly-supervised learning," *Sensors*, vol. 20, no. 9, p. 2684, 2020.

[23] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018.

[24] Z. Campilho, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Scientific Reports*, vol. 7, no. 1, 2017.

[25] S. K. Asare, F. You, and O. T. Nartey, "Efficient, ultra-facile breast cancer histopathological images classification approach utilizing deep learning optimizers," *International Journal of Computer Applications*, vol. 11, p. 9, 2020.

[26] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[27] X. Zhu, "Semi-supervised learning literature survey," 2008, http://digital.library.wisc.edu/1793/60444.

[28] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning. MIT press," 2006.

[29] X. Chang and H. Q Shen, S. Wang, "Semi-supervised feature analysis for multimedia annotation by mining label correlation in advances in knowledge discovery and data mining," *Pacific-Asia Conference on Knowledge Discovery and Data Miningpp*, vol. 8444, pp. 74–85, 2014.

[30] S. Oriane, B. Mateusz, A. Yannis, and G Guillaume, "Rethinking deep active learning: using unlabeled data at model training," 2019.

[31] Y. En, J. Sun, J. Li, X. Chang, H. Xian-hua, and A. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, pp. 1276–1288, 2019.

[32] M. Luo, L. Zhang, F. Nie, X. Chang, B. Qian, and Q. Zheng, "Adaptive semi-supervised learning with discriminative least squares regression," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, August 2017.

[33] F. Schwenker and E. Trentin, "Pattern classification and clustering: a review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.

[34] O. T. Nartey, G. Yang, J. Wu, and S. K. Asare, "Semi-supervised learning for fine-grained classification with self-training," *IEEE Access*, vol. 8, pp. 2109–2121, 2020.

[35] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2017.

[36] M. Ducoffe and F. Precioso, "Adversarial active learning for deep networks: a margin based approach," 2018.

[37] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proceedings of the Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

[38] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, vol. 1, pp. 1189–1197, Red Hook, NY, USA, December 2010.

[39] T. Pi, Z. Li, M. Zhongfei, W. Deyu, X. Jun, and Z. Yueting, "Self-paced boost learning for classification," in *Proceedings of the Twenty-fifth Aaai International Joint Conference on Artificial Intelligence*, pp. 1932–1938, New York, NY, USA, July 2016.

[40] D. Meng and Q. Zhao, "What objective does self-paced learning indeed optimize?" 2015, https://arxiv.org/abs/1511.06049.

[41] Z. L. Ma, S. Q. Liu, and D. Meng, "On convergence property of implicit self-paced objective" 2017, https://arxiv.org/abs/1703.09923.

[42] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 27-28, Seoul, Korea, October 2019.

[43] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1641–1654, 2019.

[44] Q. Xie, E. Hovy, M. H. Luong, and V. Le, "Q self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10684–10695, Columbus, Ohio, June 2020.

[45] Q. Sun, X. Li, and Y Liu, "Learning to self-train for semi-supervised few-shot classification," 2019, https://arxiv.org/abs/1906.00562.

[46] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI International Joint Conference on Artificial Intelligence*, pp. 4278–4284, San Francisco, CA, USA, Febuary 2017.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27–30, Las Vegas, NV, USA, June 2016.

[48] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, pp. 14–16, Banff, AB, Canada, April 2014.

[49] S. J. Pan and Q. Yang, "A Survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[50] A.-A. Nahid and Y. Kong, "Histopathological breast-image classification using local and frequency domains by convolutional neural network," *Information*, vol. 9, no. 1, p. 19, 2018.

[51] S. Pratiher and S. Chattoraj, "Manifold learning and stacked sparse autoencoder for robust breast cancer classification from histopathological images," 2018.

Hindawi
Computational Intelligence and Neuroscience
Volume 2020, Article ID 8837954, 12 pages
https://doi.org/10.1155/2020/8837954

Research Article

A Novel Bayesian Approach for EEG Source Localization

Vangelis P. Oikonomou and Ioannis Kompatsiaris

Information Technologies Institute, CERTH, Thessaloniki, Greece

Correspondence should be addressed to Vangelis P. Oikonomou; viknmu@iti.gr

Received 8 September 2020; Revised 28 September 2020; Accepted 15 October 2020; Published 30 October 2020

Academic Editor: Vahid Rakhshan

We propose a new method for EEG source localization. An efficient solution to this problem requires choosing an appropriate regularization term in order to constraint the original problem. In our work, we adopt the Bayesian framework to place constraints; hence, the regularization term is closely connected to the prior distribution. More specifically, we propose a new sparse prior for the localization of EEG sources. The proposed prior distribution has sparse properties favoring focal EEG sources. In order to obtain an efficient algorithm, we use the variational Bayesian (VB) framework which provides us with a tractable iterative algorithm of closed-form equations. Additionally, we provide extensions of our method in cases where we observe group structures and spatially extended EEG sources. We have performed experiments using synthetic EEG data and real EEG data from three publicly available datasets. The real EEG data are produced due to the presentation of auditory and visual stimulus. We compare the proposed method with well-known approaches of EEG source localization and the results have shown that our method presents state-of-the-art performance, especially in cases where we expect few activated brain regions. The proposed method can effectively detect EEG sources in various circumstances. Overall, the proposed sparse prior for EEG source localization results in more accurate localization of EEG sources than state-of-the-art approaches.

1. Introduction

Brain imaging techniques are important tools since they give us the ability to understand the neural mechanisms of complex human behavior in cognitive neuroscience. Also, they have clinical applications in patients with brain tumors and epilepsy where functional brain imaging is useful for neurosurgical planning and navigation [1–4]. Among various brain imaging techniques, electroencephalography (EEG) is preferable due to the low cost of EEG devices, the high temporal resolution of EEG signal, and the portability of EEG devices. The EEG is a noninvasive brain imaging technique that measures the scalp electric potentials produced by the firing of a very large number of neurons functioning inside the brain. The identification of firing neurons is very crucial since it gives us the ability to study brain dynamics in time scales of milliseconds. The identification of the electric current sources responsible for the electrical activity inside the brain based on the EEG activity recorded at the scalp (through electrodes) is one of the major problems in EEG processing. This problem is referred to as the EEG source localization [3, 4] or EEG inverse problem [3, 5].

The EEG inverse problem involves the calculation of locations and amplitudes of EEG sources given the EEG activity and the geometry and conductivity properties of the head. During the last two decades, a wide range of methods have been developed for the identification of EEG sources. These can be classified into two large groups: (a) dipole-fitting models and (b) distributed-source models. Dipole-fitting models represent the brain activity using a small number of dipoles and try to estimate the amplitudes, the orientations, and the position of a few dipoles that explain the data [4, 5]. However, these methods are sensitive to the initial guess of the number of dipoles and their initial locations. On the other hand, distributed-source methods use a large number of dipoles with fixed positions and try to estimate their amplitudes by solving a linear inverse problem [4, 5]. The EEG linear inverse problem is ill-posed since the number of EEG sources is much larger than the number of

EEG sensors. Also, the problem is becoming more difficult due to the presence of noise.

The distributed-source methods can be divided into two large families, reflecting how they deal with the dimension of time. From one side, we have methods that estimate the spatial source distribution instant by instant [3], while on the other side, we have the spatiotemporal modelling approaches [3, 4]. Both families have their advantages and disadvantages. For example, instant-by-instant (or instantaneous) methods are suitable for continuous brain scanning [3], while spatiotemporal methods are suitable for EEG sources with oscillatory activity [3]. Among the first reported instantaneous methods is that of the Minimum Norm Estimation (MNE) [6]. However, this method tends to prefer low-activity EEG sources close to the surface over strong-activity EEG sources in depth. To correct this problem, various methods have been proposed including weighted minimum norm, Loretta [7] and sLoretta [8]. The above methods need to adjust the regularization parameter through a cross-validation procedure or the L-curve method [5]. To account for the time evolution of an EEG source, authors have used spatiotemporal models [4, 9, 10]. Representative algorithms of this family are the Multiple Sparse Priors algorithm [11], the Champagne algorithm [12], and algorithms based on the Kalman Filtering [9]. Assuming that we have much larger time points than sensors, these algorithms provide us with accurate estimates on how a source evolves across time.

EEG sources could possess various properties related to the induced brain activity. For an EEG source, it is critical to know if it is focal or not [13, 14], its spatial pattern (how its neighborhood is affected) [11, 15, 16], and if the oscillatory activity is present or not across time [3, 9–11, 15]. Furthermore, a combination of EEG sources produces complex brain activity that spans across multiple spatial (and/or time) scales [15]. All these properties could be observed either in conjunction or in disjunction depending on the underlying EEG study. Furthermore, these properties are included in the overall analysis through the assumed EEG sources' model and various assumptions about the model. Clearly, the linear observation model [17, 18], the linear dynamical model (or Kalman Filters) [17, 18], and the multiple measurement vector (MMV) model [19] make different generative modelling assumptions about the underlying mechanisms that produce the EEG data.

The spatial properties of EEG sources are encoded into the linear observation model through the use of prior distributions or regularization terms. In cases where we expect localized activity (i.e., in certain types of epilepsy), a suitable assumption is to assume that EEG sources are sparse, meaning that a few of them are activated at a specific time instant. In that case, sparse prior distributions could be used [13] or regularization terms in the form of L1-norm [14, 20]. However, EEG sources can also be both sparse and spatially distributed. Based on that, many authors develop various sparsity-promoting methods by including in their method the spatially diffused property by segmenting the brain into different predefined regions [11], by using regularization terms that take into account the spatial extension of EEG sources [21], by extending the lead field matrix to multiple spatial scales [15, 16]. However, the spatial scale over which sparsity might apply remains an area of investigation.

In the present work, we propose a new framework to deal with localized (focal) activity, which can be extended in multiple spatial scales. Our contributions, with respect to the EEG source localization, are (a) a new sparse prior for the localization of EEG sources [22] and its extension to include group-sparse structures, (b) an extended (or modified) lead field matrix for the case of spatially extended EEG sources, and (c) extensive experiments using three real EEG datasets with various properties and differences between them. A preliminary version of this work has been reported in [22]. The remainder of this paper is organized as follows. In Section 2, we describe the proposed algorithmic approach for the solution of the inverse EEG problem. Then in Section 3, we present the experiments of our approach on synthetic and real EEG data. Also, a comparison of our algorithms with baseline and state-of-the-art algorithms is provided. Finally, in Section 4, we discuss our conclusions and future directions of our work.

2. Materials and Methods

2.1. Linear Observation Model. In EEG inverse problem, we desire to find the brain activity given the EEG measurements and the geometry and conductivity properties of the head. In our work, we use the distributed-source model. This means that we use a finite number of dipoles in the cortex at given locations. Hence, the potential at the scalp is a linear combination of dipoles amplitudes, represented by the following equation:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \Re^N$ is the EEG measurement vector acquired by the N electrodes, $\mathbf{x} \in \Re^{3M}$ contains the amplitudes of M dipoles along the three spatial dimensions, and $\mathbf{H} \in \Re^{N\times 3M}$ is the lead field matrix that describes the propagation of electromagnetic field from the sources to the sensors and it contains information related to the geometry and conductivity properties of the head. The vector $\mathbf{e}$ is an additive white Gaussian noise. The EEG inverse problem of the observation model of equation (1) consists of estimating the vector $\mathbf{x}$ given the data $\mathbf{y}$ and the lead field matrix $\mathbf{H}$. In the next subsection, we describe an approach for this process by using the variational Bayesian (VB) framework. More specifically, we define the hierarchical sparse prior over the amplitudes of EEG sources, the likelihood of the model, and its hyperparameters. Also, we can observe here that our instantaneous linear observation model is suitable for cases where we do not have a correlation between time samples, the noise, and sources which are nonstationary quantities, and the number of time samples is smaller than the number of sensors.

Distributed EEG source localization represents a highly ill-posed problem since the measurements are in order of 10^2 while unknowns are in order larger than 10^4. One approach to reducing the complexity of the problem is to restrict the

solutions space by reducing the number of unknowns. In this direction, two approaches are used considerably: the restriction of solutions (or EEG sources) to the cortical surface of the brain and the placement of constraints in dipole orientation [23, 24]. The above restrictions are reflected in the construction of the lead field matrix $\mathbf{H}$. In our work, we examine both the aforementioned cases.

2.2. Sparse Bayesian Learning. From a machine learning perspective, sparsity is a very helpful property since the processing is faster in a sparse representation where few coefficients reveal the information we are looking for. Hence, sparse priors help us to determine the model order in an automatic way and reduce its complexity. In addition to the above, from a brain imaging perspective, the motivation of using sparse priors is based on the localized (or focal) activity that can be observed in certain types of epilepsy and on observed sparse activations in the brain during high cognitive processing as revealed by various brain imaging techniques. In [13], sparse priors, based on a Bernoulli Laplacian prior, are used resulting in a posterior distribution where the estimators cannot be computed with close-form expressions. For this reason, the authors in [13] use the Markov Chain Monte Carlo framework.

In this work, the EEG sources $\mathbf{x}$ are treated as a random variable following a Gaussian distribution of zero mean and variance $a_i^{-1}\lambda_i^{-1}$:

$$p(\mathbf{x}|\mathbf{a};\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{x}|0,\boldsymbol{\Lambda}) = \prod_{i=1}^{3M} \mathcal{N}\left(x_i|0, a_i^{-1}\lambda_i^{-1}\right), \tag{2}$$

where $\mathcal{N}$ is the symbol for Gaussian distribution. In Sparse Bayesian Learning literature [18, 25, 26], a common approach is to assume that the covariance matrix Λ is a diagonal matrix with elements $a_i^{-1}, i = 1, \ldots, 3M$. Each parameter a_i, which controls the prior distribution of the EEG sources $\mathbf{x}$, follows a Gamma distribution, so the overall prior over all a_i is a product of Gamma distributions given by $p(\mathbf{a}) = \prod_{i=1}^{3M} \text{Gamma}(a_i; b_a, c_a)$. However, in our study, we introduce one more parameter into the distribution. More specifically, we assume that the covariance matrix Λ is a diagonal matrix with elements $a_i^{-1}\lambda_i^{-1}, i = 1, \ldots, 3M$. In our analysis, parameters λ_i are assumed to be known and deterministic quantities.

At this point, it is worth examining the marginal prior distribution of EEG source x_i by eliminating the hyperparameters a_i:

$$\begin{aligned} p(x_i;\lambda_i) &= \int p(x_i|a_i;\lambda_i)p(a_i)\mathrm{d}a_i \\ &= \int \mathcal{N}\left(x_i|0, a_i^{-1}\lambda_i^{-1}\right)\text{Gamma}(a_i; b_a, c_a)\mathrm{d}a_i \\ &\propto \left(\frac{\lambda_i}{b_a}\right)^{1/2}\left[1 + \frac{\lambda_i x_i^2}{b_a}\right]^{-\left(c_a+1/2\right)}. \end{aligned} \tag{3}$$

Equation (3) can be recognized as a Student-t distribution with zero mean, shape parameter c_a, and scale parameter b_a/λ_i. We can see that parameter λ_i controls the scale of the Student-t distribution. In addition, by adopting a procedure similar to [25], we can show that the EEG sources have the improper prior $p(x_i) \propto 1/(\lambda_i^{1/2} \cdot |x_i|)$. Now, by setting $\lambda_i \longrightarrow 1/|x_i|$, we obtain $p(x_i) \propto 1/\cdot |x_i|^{1/2}$ which can be recognized as an extremely "sparse" prior.

The overall precision (inverse variance) β of the noise follows a Gamma distribution: $p(\beta) = \text{Gamma}(\beta; b, c) = (1/(\Gamma(c)))((\beta^{(c-1)})/b^c)\exp\{-\beta/b\}$, where b and c are the scale and the shape of the Gamma distribution, respectively. We use the Gamma distribution for the noise components for two reasons: first, this distribution is conjugate to the Gaussian distribution, which helps us in the derivation of closed-form solutions, and second, it places the positivity restriction on the overall variance and the scaling parameters.

So, the overall prior over model parameters $\{\mathbf{x}, \mathbf{a}, \beta\}$ is given by $p(\mathbf{x}, \mathbf{a}, \beta; \lambda) = p(\mathbf{x}|\mathbf{a}; \lambda)\prod_{i=1}^{3M} p(a_i)p(\beta)$. The likelihood of the data is given by

$$p(\mathbf{y}|\mathbf{x}, \beta; \boldsymbol{\lambda}) = \frac{\beta^{N/2}}{(2\pi)^{N/2}} \cdot \exp\left\{-\frac{\beta}{2}(\mathbf{y} - \mathbf{H}\mathbf{x})^T(\mathbf{y} - \mathbf{H}\mathbf{x})\right\}. \tag{4}$$

To apply the VB methodology [17], we need to define an approximate posterior based on one factorization over the parameters $\{\mathbf{x}, \mathbf{a}, \beta\}$. In our study, we choose the following factorization: $q(\mathbf{x}, \mathbf{a}, \beta; \lambda) = q(\mathbf{x}|\mathbf{a}; \lambda)\prod_{i=1}^{3M} q(a_i)q(\beta)$.

Applying the VB methodology and taking into account the above factorization, the following posteriors are obtained:

$$\begin{aligned} q(\mathbf{x}) &= \mathcal{N}(\widehat{\mathbf{x}}, \mathbf{C}_{\mathbf{x}}), \\ q(\beta) &= \text{Gamma}(\beta; b', c'), \\ q(\mathbf{a}) &= \prod_{i=1}^{D} \text{Gamma}\left(a_i; b'_{a_i}, c'_{a_i}\right). \end{aligned} \tag{5}$$

The moments of each distribution are calculated by applying iteratively the following equations until convergence:

$$\mathbf{C}_{\mathbf{x}}^{(k+1)} = \left(\widehat{\beta}^{(k)}\mathbf{H}^T\mathbf{H} + \widehat{\boldsymbol{\Lambda}}^{(k+1)}\right)^{-1}, \tag{6}$$

$$\widehat{\mathbf{x}}^{(k+1)} = \left(\widehat{\beta}^{(k)}\mathbf{H}^T\mathbf{H} + \widehat{\boldsymbol{\Lambda}}^{(k+1)}\right)^{-1}\widehat{\beta}\mathbf{H}^T\mathbf{y}, \tag{7}$$

$$\frac{1}{b_{a_i}^{(k+1)'}} = \frac{\lambda_i^{(k+1)}}{2}\left(\left(\widehat{x}_i^{(k+1)}\right)^2 + \mathbf{C}_{\mathbf{x}}^{(k+1)}(i,i)\right) + \frac{1}{b_a}, \tag{8}$$

$$
\begin{aligned}
c_{a_i}^{(k+1)'} &= \frac{1}{2} + c_a, \\
q(\mathbf{a}) &= \prod_{i=1}^{D} \text{Gamma}\left(a_i; b_{a_i}', c_{a_i}'\right), \\
\frac{1}{b_\beta^{(k+1)'}} &= \frac{1}{2}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{(k+1)}\right)^T\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{(k+1)}\right) + \text{tr}\left(\mathbf{H}^T\mathbf{H}\mathbf{C}_{\mathbf{x}}^{(k+1)}\right) + \frac{1}{b}, \\
c_\beta^{(k+1)'} &= \frac{N}{2} + c, \\
\widehat{\beta}^{(k+1)} &= b_\beta^{(k+1)'} c_\beta^{(k+1)'}.
\end{aligned}
\tag{9}
$$

In the above equations, the matrix $\widehat{\Lambda}^{(k+1)}$ is a diagonal matrix with $\widehat{a}_i^{(k)} \cdot \lambda_i^{(k+1)}$ in its main diagonal. For $\lambda_i^{(k+1)}$, we follow the considerations of [27] and we set them to $1/|\widehat{x}_i^{(k)}|$. With respect to other similar approaches [25, 28], we can observe the difference in equations (7) and (8). More specifically, in our approach, the parameter b_{a_i}' is weighted by the corresponding parameter λ_i. Observe here that this parameter is affecting the scale of marginal Student-t distribution (see equation (3)).

2.3. Group-Sparse Priors. In the subsequent analysis, we assume that the EEG sources $\mathbf{x}$ have a group structure. More specifically, we define G groups of EEG sources such that the vector $\mathbf{x}_g$ contains d_g coefficients assigned to group g. Sparsity between groups can be achieved by selecting carefully the prior distribution over them. Assuming a priori independence between groups and that each group follows a Gaussian distribution with zero mean and covariance matrix $a_g^{-1}\mathbf{I}_{d_g}$, the prior over coefficients is given by

$$
p(\mathbf{x}|\mathbf{a}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{x}_g \mid \mathbf{0}_{d_g}, a_g^{-1}\mathbf{I}_{d_g}\right), \tag{10}
$$

where $\mathcal{N}$ is the symbol for Gaussian distribution. Furthermore, we assume that each parameter a_g, which controls the group sparsity of the EEG sources $\mathbf{x}$, follows a Gamma distribution, so the overall prior over all a_g is a product of Gamma distributions given by $p(\mathbf{a}) = \prod_{g=1}^{G} \text{Gamma}(a_g; b_a, c_a)$. The above hierarchical prior belongs to the family of conjugate distributions and it is well known for its sparse properties [25, 26] with respect to the groups. As before (see Section 2.2), we change the above prior by introducing one more parameter. More specifically, we assume that the prior covariance matrix is a diagonal matrix with elements $a_g^{-1}\lambda_g^{-1}$. In our analysis, parameters λ_g are assumed to be known and deterministic quantities. Now, the prior distribution of coefficients is given by

$$
p(\mathbf{x}|\mathbf{a};\boldsymbol{\lambda}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{x}_g \mid 0, a_g^{-1}\lambda_g^{-1}\mathbf{I}_{d_g}\right). \tag{11}
$$

Using the above group-sparse prior and following a similar VB procedure as that in the previous section, we can derive an iterative algorithm. More information about the derivation of the group-based algorithm can be found in [29]. Also, with respect to the above algorithm, a coefficient could potentially belong to several groups. Overlapping between groups is permitted; however, special care must be taken in order to reflect the anatomical and functional properties of the brain.

It interesting at this point to examine possible group strategies with respect to the inverse EEG problem. We can observe here that in equation (1) each dipole is represented by three components in the lead field matrix, one for each of the three spatial dimensions. So, an obvious choice of grouping is to define one group for each dipole. In that case, we have $G = M$ and $d_g = 3$ for the group-sparse prior. Another choice of grouping is to use an anatomical (or functional) template (or brain maps) to define the groups. Finally, a third option is to define the groups by using a criterion based on distances between the dipoles (i.e., dipoles in close distance are expected to behave in a similar fashion). Observe here that the first two group creation strategies are based on information related to the brain's structure, organization, and function. Also, in these cases, one dipole belongs only to one group (the groups are disjointed sets and there is no overlap between them), while in the distance-based grouping, one dipole could belong to various groups (overlapping between groups exists). In all the above cases, the structure of groups is considered known before the application of the algorithm.

2.4. Spatially Extended EEG Sources. In the above sections, we have assumed that the EEG sources are focal in nature and we examined their sparseness in the original EEG source domain. However, EEG sources could be spatially extended in cases such as in cognitive tasks or spontaneous states [3]. In this subsection, we borrow one of the general ideas from the Compress Sensing framework [30]. More specifically, we assume that EEG sources, $\mathbf{x}$, are sparse in another domain which we call it ψ-domain. In our approach, the ψ-domain could be the wavelet domain, Fourier domain, discrete cosine domain, or any other linear transformation and it is represented by the matrix $\boldsymbol{\Psi}$. The EEG sources, $\mathbf{x}$, can be written as

$$
\mathbf{x} = \boldsymbol{\Psi}\mathbf{z}, \tag{12}
$$

where $\mathbf{z}$ is a vector that contains the coefficients of EEG sources in the ψ-domain and also this vector has sparse nature due to the assumption of sources' sparseness in the ψ-domain. Now, the basic equation of our work (equation (1)) can be written in the ψ-domain as

$$
\begin{aligned}
\mathbf{y} &= \mathbf{H}\mathbf{x} + \mathbf{e} \\
&= \mathbf{H}\boldsymbol{\Psi}\mathbf{z} + \mathbf{e} \\
&= \mathbf{H}_\psi\mathbf{z} + \mathbf{e}.
\end{aligned}
\tag{13}
$$

We can observe here that the original lead field matrix has been modified by the transformation matrix Ψ, $\mathbf{H}_\psi = \mathbf{H}\Psi$. Using one of the previous algorithms (or any other sparsity induced algorithm), we can find the coefficients $\mathbf{z}$, and finally, the EEG sources can be obtained by using equation (12).

The choice of ψ-domain (which is reflected in the structure of matrix Ψ) is crucial for the properties of

original EEG sources, $\mathbf{x}$. Also, this choice must incorporate some prior knowledge about the original EEG sources. Observe here that the EEG sources are positioned on a grid in 3D space; hence, direct use of wavelet transform or Fourier transform is not an easy task. Furthermore, interpretation of the results from a neurophysiological viewpoint is more difficult. Since our goal is to find spatially extended EEG sources, we adopt a local spatial smoothing kernel [16]. More specifically, for i-th EEG source, we define

$$\psi_{ij} = \begin{cases} 1, & \text{if } i = j, \\ \exp\left(-r \cdot d_{ij}^2\right), & \text{if } i \neq j, \end{cases} \tag{14}$$

where $d_{ij}, i = 1, \ldots, N, j = 1, \ldots, N$ is the spatial distance between the i-th and j-th EEG sources, while r is a parameter that controls the extension of spatial smoothness between individual EEG sources. In our work, the parameter r is assumed to be known; however, we can estimate it by using a cross-validation approach or methods based on multiple kernel learning [31, 32]. Looking at equation (12), we can verify that the original EEG sources are spatially extended due to $\Psi\mathbf{z}$ and the properties of the vector $\mathbf{z}$ (sparsity) and the matrix Ψ (spatially extended).

Concluding this section, we want to mention that three approaches, using the Bayesian framework, are provided. The first approach (we call it *Fan*) is described in Section 2.1 and it presents the backbone of our overall methodology. This method is suitable for finding focal EEG sources due to its sparse properties. The second approach (we call it *FanGr*) is an extension of *Fan* approach. The main characteristic of this method is that now we can define groups over EEG sources. Finally, the third approach (we call it *FanSmooth*) is similar to the first approach but with one critical difference in the lead field matrix. In this last approach, we use a modified lead field matrix using ideas from CS framework.

3. Experiments and Results

In this section, we present our experiments with the corresponding results using synthetic EEG data and real EEG data from three EEG experiments. The real EEG data are produced due to the presentation of auditory and visual stimulus on the participants. In all our experiments, we have used the FieldTrip toolbox [33] to preprocess the EEG data and to construct the lead field matrices. In our study, we adopted two approaches for the construction of lead field matrices, the cortical-based approach, and the volumetric-based approach.

3.1. Experiments Using Synthetic EEG Data. Synthetic data with few pointwise source activations (see equation (1)) were generated using realistic head models with electrodes placed according to the 10–10 international system of electrode placement. In our study, we investigate two cases of, with respect to the number of channels, 128 channels and 256 channels.

3.1.1. Activations. In our work, we investigated two different kinds of activations: (1) single dipole activations, and (2) multiple dipole activations. The first case represents a situation where one dipole is activated among many, and the second case represents a situation where many dipoles (possibly distant) are activated. The amplitudes of active EEG sources were samples from a Gaussian distribution with zero mean and variance one. Finally, with respect to EEG measurements, we examine two cases: noise-free measurements and noisy measurements. In noisy measurements, we added white Gaussian noise and the signal-to-noise ratio (SNR) was defined to 60 dB.

3.1.2. Lead Field Matrix. With respect to the lead field matrix, we examined two cases for its construction: the cortical-based case and the volumetric-based case. *Cortical based*: in this case, the dipoles are placed on a spatial grid covering the cortical surface. The positions and orientations of dipoles are fixed. In addition, orientations are normals to the cortical surface [13, 24]. Finally, from the perspective of neurophysiology, the source space is the cortex (i.e., we assume that the observed electrical activity is produced by a specific brain structure). The number of dipoles was 5124; hence, the resulting lead field matrix is $\mathbf{H} \in \mathfrak{R}^{128\times5124}$ or $\mathbf{H} \in \mathfrak{R}^{256\times5124}$. *Volumetric (or grid) based*: in this case, the dipoles are placed on a spatial grid covering the entire brain. Also, the positions of dipoles are fixed but the orientations are free. In addition, the source space includes the cortex, subcortical structures, and the cerebellum. The grid resolution was set to 1 cm resulting in 2020 dipoles; hence, the resulting lead field matrix is $\mathbf{H} \in \mathfrak{R}^{128\times6060}$ or $\mathbf{H} \in \mathfrak{R}^{256\times6060}$. Overall, in this set of experiments, we examine configurations of inverse EEG problems with respect to the number of channels, the type of lead field matrix, the presence (or not) of noise, and the type of activations. Each configuration is repeated 50 times in order to obtain averaged results with respect to the performance of each method.

3.1.3. Performance Measures. In order to evaluate the performance of an algorithm, we adopt the following measures. *Reconstruction error*: we use the reconstruction error between the true EEG sources, $\mathbf{x}_{\text{true}}$, and the estimated EEG sources, $\mathbf{x}_{\text{est}}$, given by $\|\mathbf{x}_{\text{est}} - \mathbf{x}_{\text{true}}\|_2^2 / \|\mathbf{x}_{\text{true}}\|_2^2$. This measure will determine whether the algorithm recovers the source energy. *Localization error* [20]: we use the Euclidean distance between the simulated source and the maximum of the estimated activity within the sphere neighboring the simulated source. This measure will determine whether the algorithm is able to find the point of the simulated source. In our study, the neighbor was set to 25 mm [20]. A' *metric* [16]: this metric is computed as $A' = ((H_R - F_R)/2) + (1/2)$, where H_R is the hit rate and F_R is the false positive rate. This measure estimates the area under the Receiver Operator Characteristic (ROC) curve and it is related to the detection accuracy of the algorithm (if the area under the ROC is large, then the hit rate is high compared to the false positive rate). In order to define the hit rates, we follow a similar procedure to that of [16], where we included in the calculation of hit

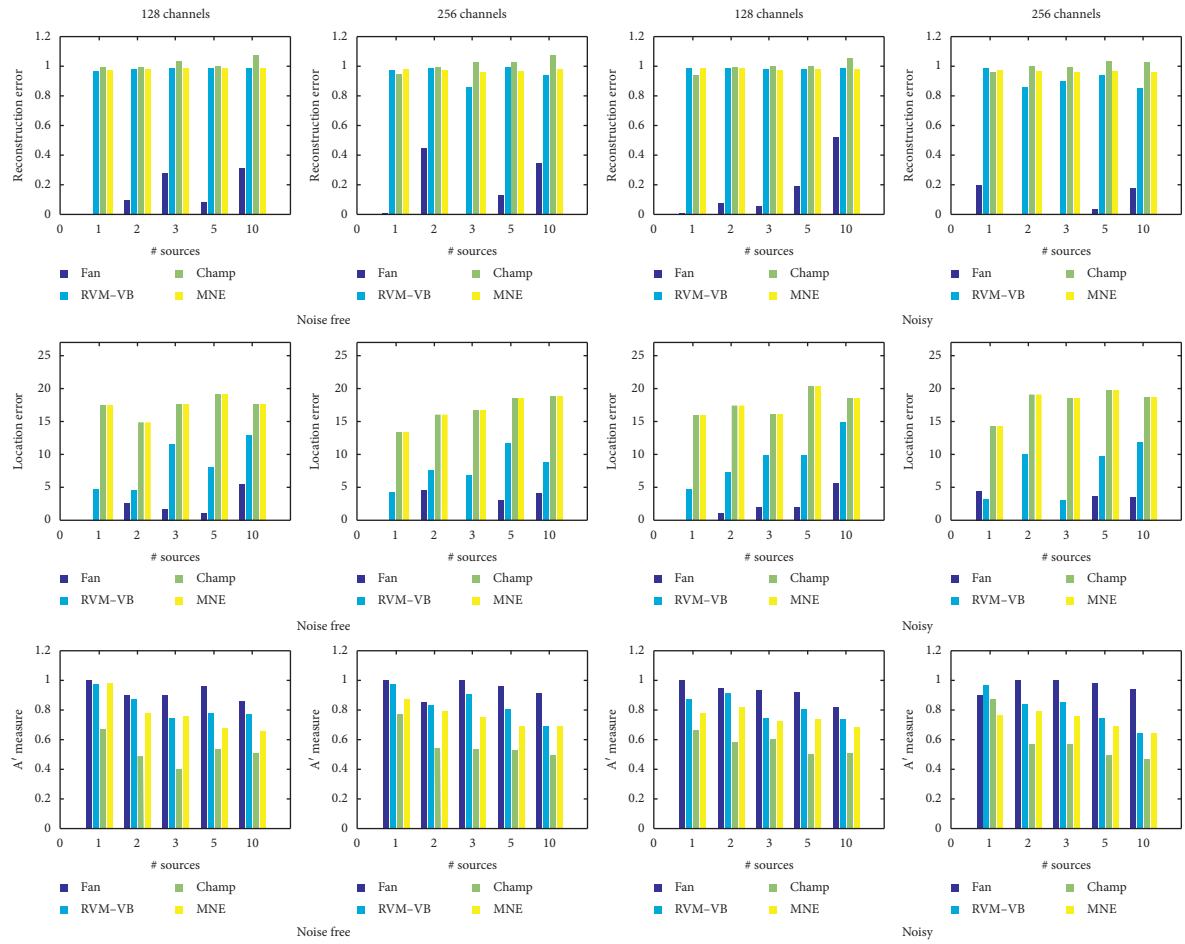


Figure 1: Obtained performance measures in the case of cortical-based lead field matrix.

rates voxels that are at least 0.1% of the maximum activation of the localization result. Finally, we compared our methods with the following approaches: (a) the Minimum Norm Estimator (*MNE*) [4, 6], a classical approach for the EEG inverse problem, (b) the Relevance Vector Machines using the VB approach (*RVM-VB*) [28], and (c) the plain Champagne (*Champ*) [4, 12] using the available code from the NUTMEG toolbox [34].

3.1.4. Results on Synthetic EEG Data. In Figure 1, we provide the obtained results when a cortical-based lead field matrix is used with respect to all performance measures. The results are shown with respect to the measures, the number of active EEG sources, the number of channels, and the presence (or not) of noise. We can see that the proposed approach presents the best performance compared to other methods. More specifically, the proposed method presents the smallest reconstruction and location error and the highest value for A' metric. This is observed in all cases irrespective of the number of active EEG sources or the number of channels or to the presence of noise. Additionally, in Figure 2, we present the obtained results when the volumetric-based lead field matrix is used. In this set of experiments, we use, also, the group version of our method since one dipole can be considered as a group of three elementary dipoles (one for each of three spatial dimensions). We observe that both versions of our approach present better performance (in terms of reconstruction error, location error, and A' metric) than the other methods. Also, we can see that, for the majority of activation profiles, the adoption of grouping structures increases the performance of our analysis, especially when we have multiple activations. Clearly, the proposed approach is able to reconstruct more accurately the spatial pattern of EEG sources without introducing error in the location of EEG source(s) resulting in high detection accuracy.

3.2. Experiments Using Real EEG Data. In this section, we provide our results from experiments using real EEG data from three EEG datasets. The EEG experiments were designed to study brain responses with respect to auditory and visual stimuli. Furthermore, in this section, we include in our analysis the *FanSmooth* ($r = 0.05$) method. The value for spatial smoothness r has been determined after the empirical evaluation of obtained brain maps.

3.2.1. Experiments Using Auditory EEG Data. In this section, we perform experiments using EEG data that corresponds to

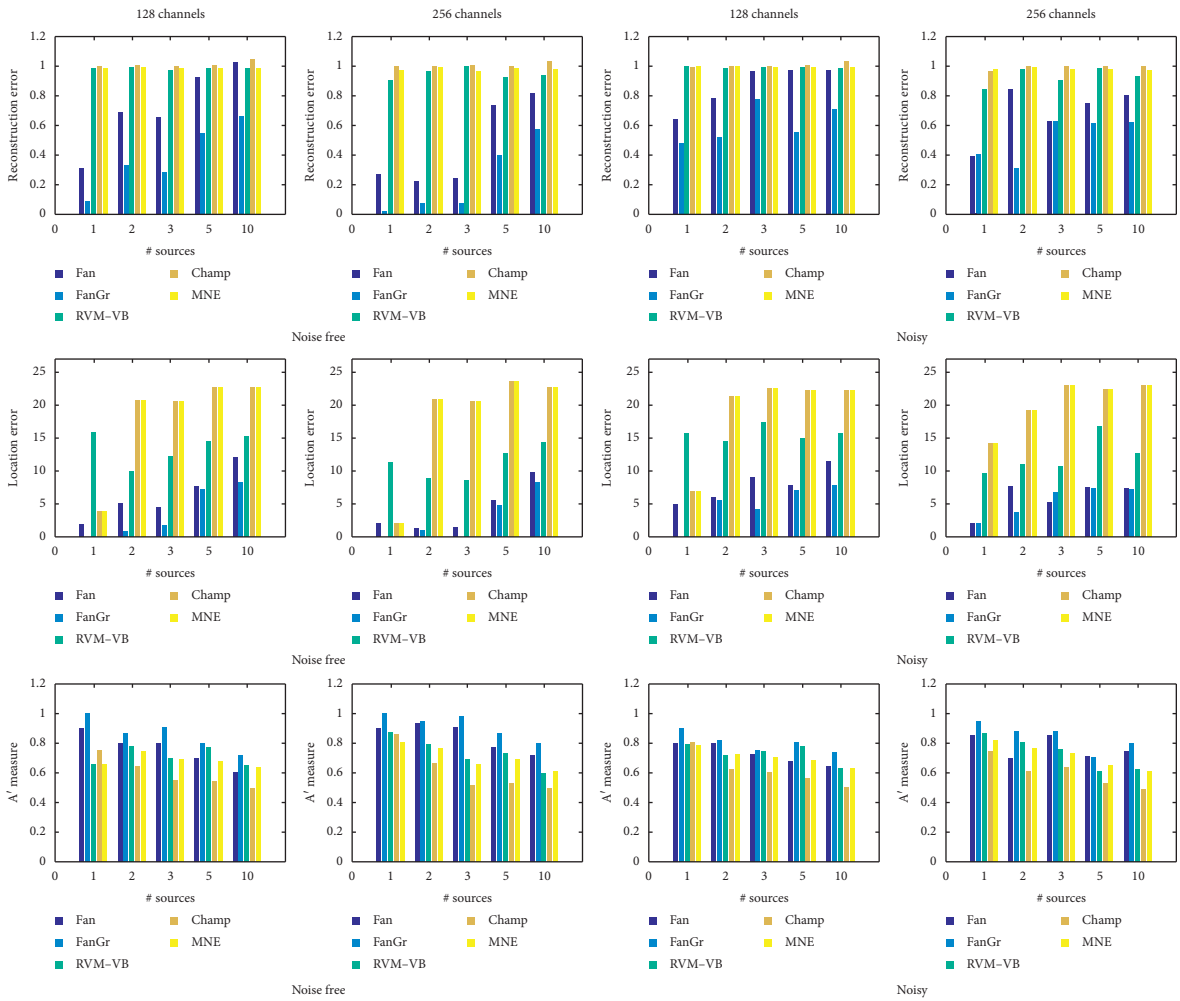


Figure 2: Obtained performance measures in the case of volumetric-based lead field matrix.

an auditory oddball paradigm and they can be downloaded from the homepage of the FieldTrip toolbox2. The raw EEG data consist in 600 trials. The duration of each trial was 2 secs, 1sec of EEG data preceding the acoustic stimulus, and 1sec of EEG data following the stimulus. The EEG activity was recorded using 128 channels at 1000 Hz. The EEG trials were band-pass filtered at 1–40 Hz and downsampled at 250 Hz. A realistic head model was used based on cortical surface approach. The number of dipoles was 5124; hence, the resulting lead field matrix is $\mathbf{H} \in \mathfrak{R}^{128\times5124}$. ERPs were formed by averaging over all trials. In this experiment, brain sources are detected by algorithms for the time point that corresponds to the peak of the electrical activity in the frontal-central scalp in the time range between 100 ms and 200 ms.

The estimated brain activity using the aforementioned methods is shown in Figure 3. The *Fan*, *FanSmooth*, *RVM-VB*, and *Champ* methods present activations in the temporal lobe, as expected in auditory experiments. However, the *Fan*, the *FanSmooth*, and the *Champ* methods provide activations on both hemispheres of the temporal lobe, while the *RVM-VB* method provides activations only to the right temporal lobe. The *MNE* method does not show activation in the temporal lobe. In addition to the above, we observe that all methods, besides *Champ*, present activations in the right frontal lobe. This type of activation is not unusual in auditory experiments, especially when deviant tones are involved [35, 36].

3.2.2. Experiments Using Visual (Facial) Evoked Potentials EEG Data. The EEG data used in this section is part of the Multimodal Face Dataset available in the SPM software 3. This dataset was acquired from a face perception study in which the subject had to judge the symmetry of a mixed set of faces and scrambled faces. More details about the dataset can be found in [37]. The EEG acquisition system was a 128-channel ActiveTwo Biosemi system with a sampling frequency equal to 2048 Hz. The data were downsampled to 256 Hz, and after artifact rejection, the 309 epochs were averaged and low-pass filtered at 20 Hz. A realistic head model was used based on cortical surface approach. The number of dipoles was 5124; hence, the resulting lead field matrix is $\mathbf{H} \in \mathfrak{R}^{128\times5124}$.

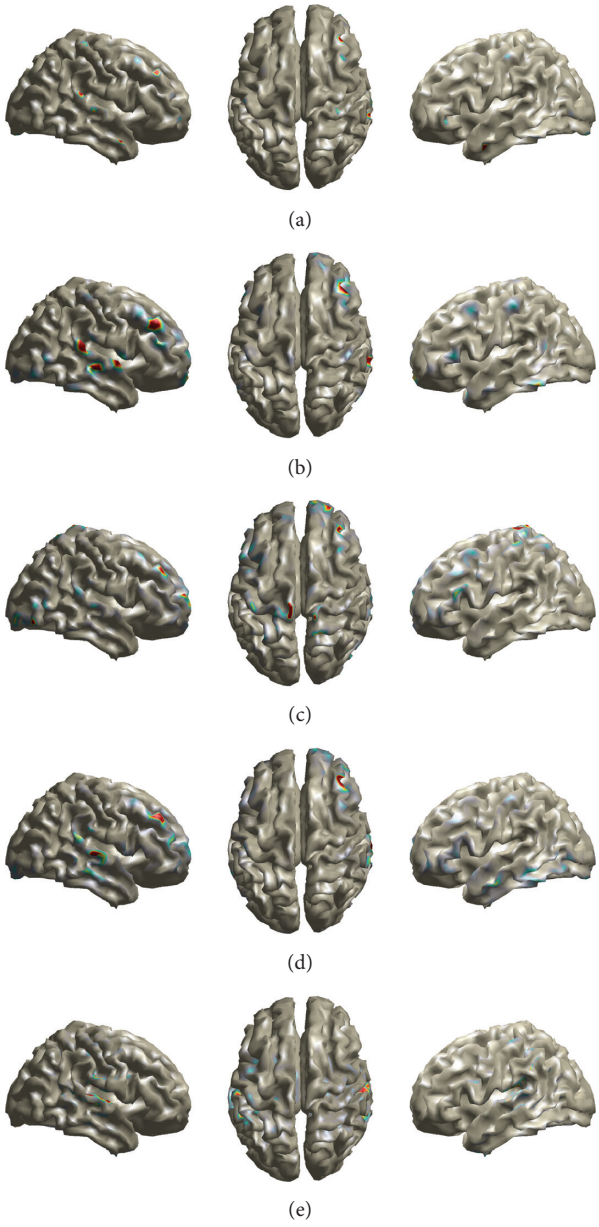


Figure 3: Brain maps showing EEG sources in the case of auditory EEG data. (a) Fan. (b) Fan smooth. (c) MNE. (d) RVM-VB. (e) Champ.

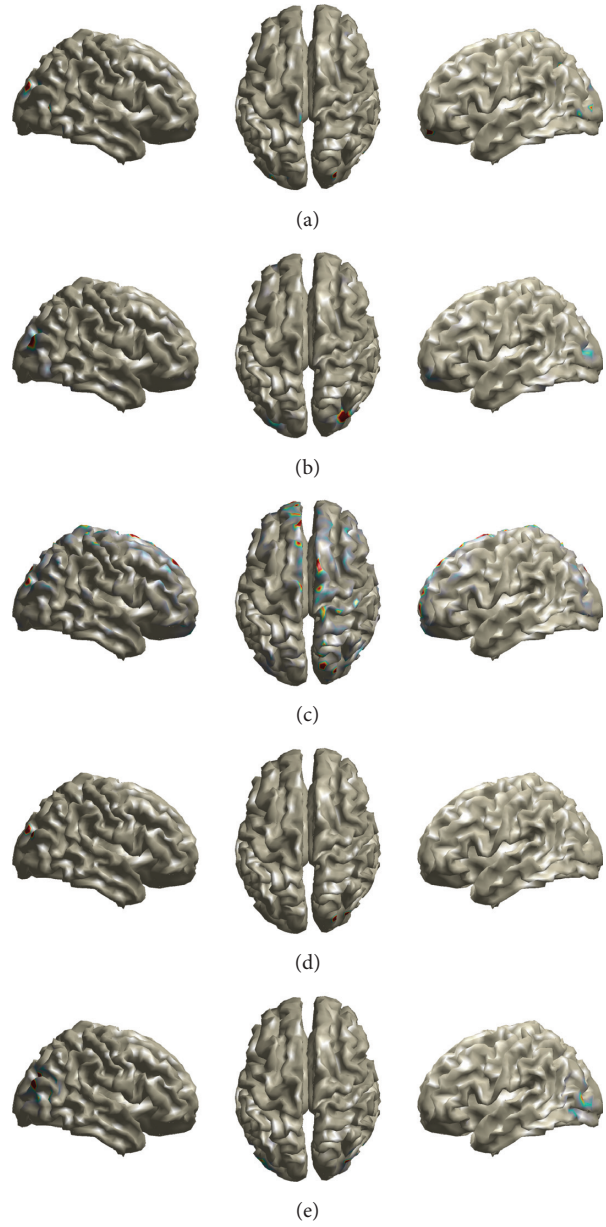


Figure 4: Brain maps showing EEG sources in the case of Visual (Faces) EEG data. (a) Fan. (b) Fan smooth. (c) MNE. (d) RVM-VB. (e) Champ.

The estimated activities from all methods are shown in Figure 4 (at 100 ms). Careful inspection of these images reveals that all methods present their primary activations on the occipital lobe as expected in this kind of experiment. However, we can also observe substantial differences with respect to the type of activation. More specifically, the *RVM-VB* and the *Fan* methods present the most compact activated area compared to other methods. Additionally, the *Fan*, the *FanSmooth*, and the *Champ* methods present bilateral activation on the occipital lobe, while the *RVM-VB* and the *MNE* methods present activation only to the right occipital lobe. Furthermore, the *Fan* and the *MNE* methods present secondary activations on the frontal lobe. In addition to that, the *MNE* method presents activations to the Supplementary Motor Area.

3.2.3. Experiments Using Steady-State Visual Evoked Potentials EEG Data. In this subsection, the EEG data corresponds to a Steady-State Visual Evoked Potentials (SSVEP) Brain-Computer Interface (BCI) paradigm [38]. In this dataset, 40-target visual stimuli were presented on a 23.6 in LCD monitor. Thirty-five healthy subjects with normal or corrected-to-normal vision participated in this study. EEG data were recorded with 64 electrodes according to an extended 10–20 system in order to record whole-head EEG. Data epochs were extracted according to event triggers generated by the stimulus program. All data epochs were downsampled to 250 Hz. The EEG data have been band-pass (zero phases) filtered from 4 Hz to 90 Hz with an infinite impulse response (IIR) filter (by using the *filtfilt* function in MATLAB). From this dataset for our

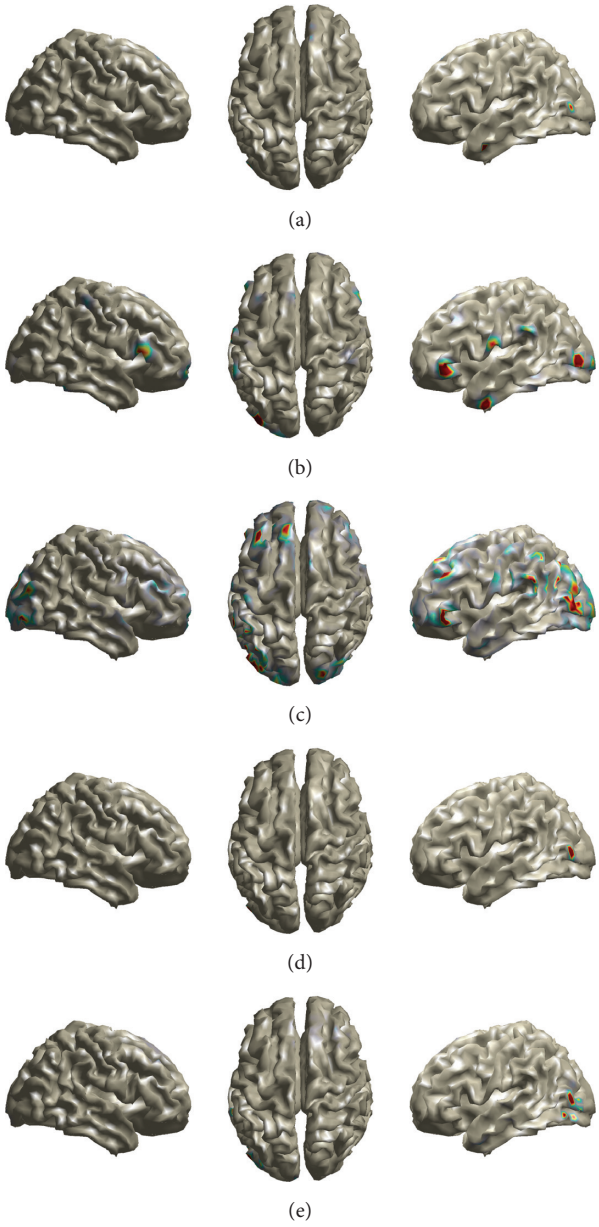


Figure 5: Brain maps showing EEG sources in the case of SSVEP EEG data. (a) Fan. (b) Fan smooth. (c) MNE. (d) RVM-VB. (e) Champ.

analysis, we have used the EEG trials from the first subject which are corresponding to the first target.

In this experiment, brain sources are detected by calculating the average scalp electrical activity between 1 sec and 4 sec. The estimated brain activity for all algorithms is shown in Figure 5. We can observe that all algorithms provide activated areas in the left part of the occipital lobe. In addition to that, the *MNE* methods provide also activations on the right part of the occipital lobe. Furthermore, we can observe activations on the frontal lobe from *FanSmooth* and *MNE* methods, while the *Fan* and *FanSmooth* methods provide an additional activation on the temporal lobe.

Concluding this section with real EEG data, it is worth providing a qualitative comparison between the methods

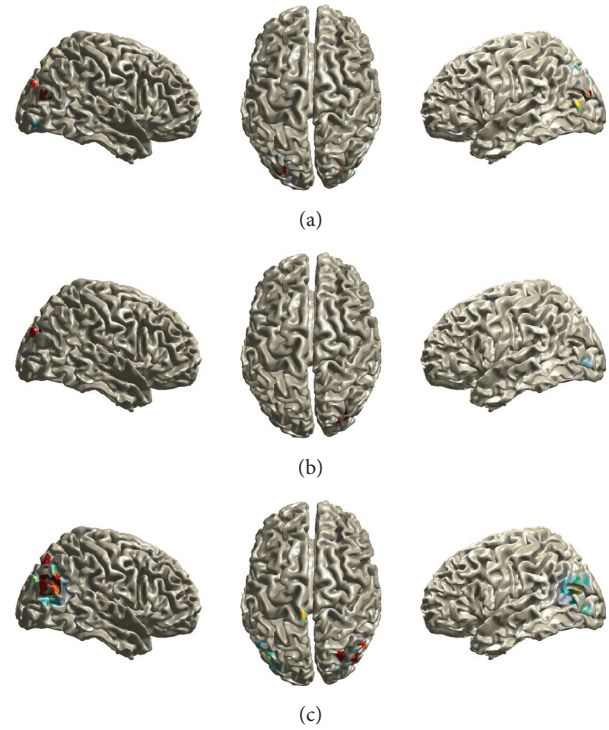


Figure 6: Brain maps (projected on the cortical surface) showing EEG sources in the case of Visual (Faces) EEG data when a volumetric-based lead field matrix is used. (a) Fan. (b) FanGr. (c) Champ.

and their properties. The *Fan* algorithm provides the most compact activated areas compared to other methods due to their inherent characteristic of sparseness. This observation is justified by observing the results when real EEG data are used as well as the "theoretical" implications of equation (3). On the other side, the *FanSmooth* algorithm provides a spatially extended activated area. Between these two extreme cases lie the *RVM-VB* algorithm and the *Champ* algorithm. However, this was expected due to the fact that (1) the *RVM-VB* algorithm and the *Champ* algorithm use a similar prior for EEG sources, which does not encourage sparsest solutions than our proposed prior, and (2) the basic version of them cannot handle spatially extended sources.

3.2.4. Volumetric Lead Field Matrix. In this section, we provide experiments using the Faces EEG data. However, we have used a volumetric lead field matrix where the dipoles are placed on a spatial grid covering the entire brain. The grid resolution was set to 1 cm resulting in 2020 dipoles; hence, the resulting lead field matrix is $\mathbf{H} \in \Re^{128\times(3\times2020)}$. Our goal in these experiments is to explore the behavior of our algorithms when groups of elementary dipoles are present. We perform a comparison between *FanGr*, *Fan*, and *Champ* algorithms. The *FanGr* algorithm is an extension of *Fan* algorithm when we want to utilize groups of dipoles, while we have used *Champ* algorithm as a baseline algorithm for comparative purposes.

In Figure 6, we provide the estimated activity of the aforementioned algorithms for the Faces EEG data. The

preprocessing steps of EEG data are described in Section 3.2.2. We can observe that all algorithms provide activation in the occipital lobe as expected. However, we can observe differences in the pattern of activations. The activated area is larger in the *Champ* algorithm, followed by *Fan* algorithm, and, lastly, the *FanGr* algorithm provides the smallest activated area in the occipital lobe. We can, also, observe that the strength of activation is stronger in the left part of the occipital lobe for the *FanGr* and *Champ* algorithms, while the *Fan* algorithm presents strong activations on both parts of the occipital lobe. In addition to the above, we can observe that the *Champ* algorithm provides a secondary activation in the parietal lobe which cannot be justified by the type of experiments and the results that we obtained by all other algorithms and the two lead field matrices; hence, we assume that this activation is a spurious one. Concluding this section, we want to mention that both types of lead field matrices do not affect considerably the obtained results, irrelevant to the method that it was used to solve the inverse EEG problem. However, this observation is also affected by the type of EEG experiment.

4. Conclusions

In this work, we proposed a new algorithm (and its gradual extensions) to solve the EEG inverse problem. In this type of inverse problems, crucial part has the regularization term. In order to regularize the EEG inverse problem, we adopt the Bayesian approach; hence, regularizations are incorporated into the overall procedure in terms of prior distributions. Furthermore, we proposed new sparse priors for the modelling of EEG sources. The main contribution of these priors is that now we are able to examine the notion of sparseness in EEG source modelling, using structures of groups. Additionally, the basic idea of CS framework was used to provide us with modified lead field matrices specialized in modelling spatially extended EEG sources. Under the Bayesian formulation, the posterior distribution in our problem was intractable and to figure out this problem, we adopted the VB framework. The proposed Bayesian methods have been tested using head models with different geometries. The obtained results, using synthetic and real EEG data, show the merits of our methods in the estimation of EEG sources.

In the future, our research will be focused on accurate modelling of the head's properties and spatiotemporal extensions of our method with applications in the BCI domain [39–41]. More specifically, we intend to combine head models with different head geometries and tissue conductivities by adopting the multikernel learning methodology. The multikernel approach could lead us to the simultaneous estimation of the extended (or composite) lead field matrix and the EEG sources in an iterative fashion. Furthermore, spatiotemporal versions of our model based on the MMV model [1, 2, 19] could be devised in order to study EEG microstates [42] in BCI domain. In addition to the above, borrowing ideas from image superresolution [43], we could provide brain imaging techniques with increased spatial resolution. Finally, the EEG source localization has close connections with CS theory [30, 44]. However, typical approaches on the construction of lead field matrix do not produce a sensing matrix with the two basic properties of CS theory, the incoherence and the restricted isometry property. It is important to investigate procedures that could provide us with a lead field matrix that possesses these two properties.

Data Availability

Auditory EEG data had been found at http://www.fieldtriptoolbox.org. Face Perception EEG data had been found at https://www.fil.ion.ucl.ac.uk/spm/data. SSVEP EEG data are freely available from http://bci.med.tsinghua.edu.cn/download.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was a part of project MAMEM that had received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement no. 644780 and project NeuroMkt that had been cofinanced by the European Regional Development Fund of the European Union and Greek National Funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE INNOVATE (Project code T2EDK-03661).

References

[1] S. Liu, R. Cao, Y. Huang, T. Ouypornkochagorn, and J. Ji, "Time sequence learning for electrical impedance tomography using bayesian spatiotemporal priors," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6045–6057, 2020.

[2] S. Liu, Y. Huang, H. Wu, C. Tan, and J. Jia, "Efficient multitask structure-aware sparse bayesian learning for frequency-difference electrical impedance tomography," *IEEE Transactions on Industrial Informatics*, vol. 20, p. 1, 2020.

[3] B. He, L. Yang, C. Wilke, and H. Yuan, "Electrophysiological imaging of brain activity and connectivity—challenges and opportunities," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 1918–1931, 2011.

[4] K. Sekihara and S. S. Nagarajan, *Electromagnetic Brain Imaging - A Bayesian Perspective*, Springer, Berlin, Germany, 2015.

[5] R. Grech, T. Cassar, J. Muscat et al., "Review on solving the inverse problem in EEG source analysis," *Journal of Neuro-Engineering and Rehabilitation*, vol. 5, 2008.

[6] M. S. Hämäläinen and R. J. Ilmoniemi, "Interpreting magnetic fields of the brain: minimum norm estimates," *Medical & Biological Engineering & Computing*, vol. 32, no. 1, pp. 35–42, 1994.

[7] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain," *International Journal of Psychophysiology*, vol. 18, no. 1, pp. 49–65, 1994.

[8] R. Pascual-Marqui, "Standardized low resolution brain electromagnetic tomography (sloreta): technical details," *Methods and findings in experimental and clinical pharmacology*, vol. 24, no. 2, pp. 5–12, 2002.

[9] E. Pirondini, B. Babadi, G. Obregon-Henao et al., "Computationally efficient algorithms for sparse, dynamic solutions to the EEG source localization problem," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1359–1372, 2018.

[10] F. Costa, H. Batatia, T. Oberlin, C. D'Giano, and J.-Y. Tourneret, "Bayesian eeg source localization using a structured sparsity prior," *NeuroImage*, vol. 144, pp. 142–152, 2017.

[11] K. Friston, L. Harrison, J. Daunizeau et al., "Multiple sparse priors for the m/eeg inverse problem," *NeuroImage*, vol. 39, no. 3, pp. 1104–1120, 2008.

[12] D. P. Wipf, J. P. Owen, H. T. Attias, K. Sekihara, and S. S. Nagarajan, "Robust bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using meg," *NeuroImage*, vol. 49, no. 1, pp. 641–655, 2010.

[13] F. Costa, H. Batatia, L. Chaari, and J.-Y. Tourneret, "Sparse EEG source localization using Bernoulli laplacian priors," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2888–2898, 2015.

[14] J. C. Bore, C. Yi, P. Li et al., "Sparse EEG source localization using LAPPS: least absolute l-P (0)," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 7, pp. 1927–1939, 2019.

[15] C. Cai, K. Sekihara, and S. S. Nagarajan, "Hierarchical multiscale bayesian algorithm for robust MEG/EEG source reconstruction," *NeuroImage*, vol. 183, pp. 698–715, 2018.

[16] C. Cai, M. Diwakar, D. Chen, K. Sekihara, and S. S. Nagarajan, "Robust empirical bayesian reconstruction of distributed sources for electromagnetic brain imaging," *IEEE Transactions on Medical Imaging*, vol. 39, p. 1, 2019.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, Berlin, Germany, 2007.

[18] W. Wu, S. Nagarajan, and Z. Chen, "Bayesian Machine Learning: EEG\/MEG signal processing measurements," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 14–36, 2016.

[19] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, 2011.

[20] P. Xu, Y. Tian, H. Chen, and D. Yao, "Lp norm iterative sparse solution for eeg source localization," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 400–409, 2007.

[21] S. Castaño-Candamil, J. Höhne, J.-D. Martínez-Vargas, X.-W. An, G. Castellanos-Domínguez, and S. Haufe, "Solving the EEG inverse problem based on space-time-frequency structured sparsity constraints," *NeuroImage*, vol. 118, pp. 598–612, 2015.

[22] V. P. Oikonomou and I. Kompatsiaris, "Sparse EEG source localization under the variational bayesian framework," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 598–602, Athens, Greece, October 2019.

[23] T. Auranen, A. Nummenmaa, M. S. Hämäläinen et al., "Bayesian inverse analysis of neuromagnetic data using cortically constrained multiple dipoles," *Human Brain Mapping*, vol. 28, no. 10, pp. 979–994, 2007.

[24] D. E. Hyde, F. H. Duffy, and S. K. Warfield, "Voxel-based dipole orientation constraints for distributed current estimation," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 7, pp. 2028–2040, 2014.

[25] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[26] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MS, USA, 2012.

[27] G. Deng, "Iterative learning algorithms for linear Gaussian observation models," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2286–2297, 2004.

[28] C. M. Bishop and M. Tipping, "Variational relevance vector machines," in *Proceedings of the Sixteenth Conference on Uncertainty In Artificial Intelligence, UAI'00*, pp. 46–53, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, September 2000.

[29] V. P. Oikonomou, S. Nikolopoulos, and I. Kompatsiaris, "A novel compressive sensing scheme under the variational bayesian framework," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO 2019)*, pp. 1–4, A Coruña, Spain, September 2019.

[30] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[31] V. P. Oikonomou and K. Blekas, "An adaptive regression mixture model for fmri cluster analysis," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 649–659, 2013.

[32] V. P. Oikonomou, S. Nikolopoulos, and I. Kompatsiaris, "A bayesian multiple kernel learning algorithm for ssvep bci detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 5, pp. 1990–2001, 2019.

[33] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational Intelligence and Neuroscience*, vol. 2011, 2011.

[34] S. S. Dalal, J. Zumer, V. Agrawal, K. Hild, K. Sekihara, and S. Nagarajan, "Nutmeg: a neuromagnetic source reconstruction toolbox," *Neurology and Clinical Neurophysiology: NCN*, vol. 2004, p. 52, 2004.

[35] S. A. Huettel and G. McCarthy, "What is odd in the oddball task?" *Neuropsychologia*, vol. 42, no. 3, pp. 379–386, 2004.

[36] M. I. Garrido, J. M. Kilner, K. E. Stephan, and K. J. Friston, "The mismatch negativity: a review of underlying mechanisms," *Clinical Neurophysiology*, vol. 120, no. 3, pp. 453–463, 2009.

[37] R. N. Henson, Y. Goshen-Gottstein, T. Ganel, L. J. Otten, A. Quayle, and M. Rugg, "Electrophysiological and haemodynamic correlates of face perception, recognition and priming," *Cerebral Cortex*, vol. 13, pp. 793–805, 2003.

[38] Y. Wang, X. Chen, X. Gao, and S. Gao, "A benchmark dataset for SSVEP-based brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1746–1752, 2017.

[39] S. Nikolopoulos, P. C. Petrantonakis, K. Georgiadis et al., "A multimodal dataset for authoring and editing multimedia content: the MAMEM project," *Data in Brief*, vol. 15, pp. 1048–1056, 2017.

[40] V. P. Oikonomou, G. Liaros, K. Georgiadis et al., "Comparative evaluation of state-of-the-art algorithms for ssvep-based bcis," *ArXiv*, vol. 162, Article ID 00904, 2016.

[41] V. Shenoy Handiru, A. P. Vinod, and C. Guan, "Eeg source imaging of movement decoding: the state of the art and future directions," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 4, no. 2, pp. 14–23, 2018.

[42] C. M. Michel and T. Koenig, "Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review," *NeuroImage*, vol. 180, pp. 577–593, 2018.

[43] J. Yang and T. Huang, "Image super-resolution: historical overviewand future challenges," in *Super-Resolution Imaging*, P. Milanfar, Ed., pp. 3–35, CRC Press, Boca Raton, FL, USA, 2010.

[44] A. Hashemi and S. Haufe, "Improving eeg source localization through spatio-temporal sparse bayesian learning," in *Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1935–1939, Eternal City, Italy, September 2018.