Generalization Performance of Deep Learning Models in Neurodegenerative Disease Classification

Ekin Yagis, Alba G. Seco De Herrera, Luca Citi
Computer Science and Electrical Engineering (CSEE)
University of Essex
Colchester, United Kingdom
{e.yagis,alba.garcia,lciti}@essex.ac.uk

***Abstract*—Over the past decade, machine learning gained considerable attention from the scientific community and has progressed rapidly as a result. Given its ability to detect subtle and complicated patterns, deep learning (DL) has been utilized widely in neuroimaging studies for medical data analysis and automated diagnostics with varying degrees of success. In this paper, we question the remarkable accuracies of the best performing models by assessing generalization performance of the state-of-the-art convolutional neural network (CNN) models on the classification of two most common neurodegenerative diseases, namely Alzheimer's Disease (AD) and Parkinson's Disease (PD) using MRI. We demonstrate the impact of the data division strategy on the model performances by comparing the results derived from two different split approaches. We first evaluated the performance of the CNN models by dividing the dataset at the subject level in which all of the MRI slices of a patient are put into either training or test set. We then observed that pooling together all slices prior to applying cross-validation, as erroneously done in a number of previous studies, leads to inflated accuracies by as much as 26% for the classification of the diseases.**

***Index Terms*—Parkinson's Disease, Alzheimer's Disease, Deep Learning, Transfer Learning, VGG16, Resnet50, MRI, Neuroimaging**

I. Introduction

Deep learning (DL) models have attracted a great deal of research interest in medical imaging due to their advantages and successes in various fields such as image and speech recognition, automation, security, computer-aided diagnosis (CAD), just to name a few. In particular, medical image analysis using DL opened a new door into CAD. In recent years, convolutional neural networks (CNNs) have been used to detect and classify a range of diseases from cancer to neurological disorders [2]–[5].

The CNN models used in these studies are mostly utilized on well-known big datasets such as ImageNet [6] and MNIST [7]. A sample CNN architecture used in medical image classification can be seen in Figure 1. Model training and testing are generally done by splitting the dataset into three subsets: training, validation, and test. Training and validation are used to learn parameters and decide whether training is complete, whereas test data are used to evaluate model performance on new previously unseen data. However, CNN models may not perform well when presented with the new data as well as previously believed [8]. A recent study in computer vision has indicated that the true generalization performance of even classic CIFAR-10 photograph classification CNNs to new data are questionable and lower than previous results [9]. In domains such as disease detection, that kind of mismatch can cause serious problems as the researchers could design models which perform well on the specific test set but are incapable of generalizing, and fail when new data are presented [10].

It has been long known that having an appropriate data division is crucial to achieve a generalization performance [11], [12]. There are various statistical sampling techniques such as simple random sampling [13], deterministic methods [14], DUPLEX [15], and stratified sampling [16] which may be used in different types of data to decrease the variance of the model performance.

In most image classification applications, the data are randomly divided into training, validation and test sets. To measure the model's ability to adapt properly to new, previously unseen data, the ideal test set should be the reflection of the data that could be encountered elsewhere.

However, in medical image classification, the accuracy on a test set which is randomly sampled from the data may not reflect the model's performance on new, previously unseen data and may create a major bias which can be explained as data leakage [17], [18]. Generally, data leakage is a phenomenon caused by the presence of the same data both in the training and testing processes. A more subtle version of this problem is when the test data are disjoint from the training data but come from a distribution that is more similar to that of the training set than one would expect from new data [19], [20]. In 3D medical imaging such as MRI or CT, dividing the overall data randomly causing slices or patches from the same patient to be in both training and test sets and leads to a biased assessment.

In this work, we assessed the generalization performance of the networks on the classification of the two most common neurological disorders: Parkinson's Disease (PD) and Alzheimer's Disease (AD). The contributions of this paper are as follows:

- We proposed a framework for PD and AD classification using CNNs and MR images;
- We utilized two state of the art convolutional neural network models together with a smart data selection

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

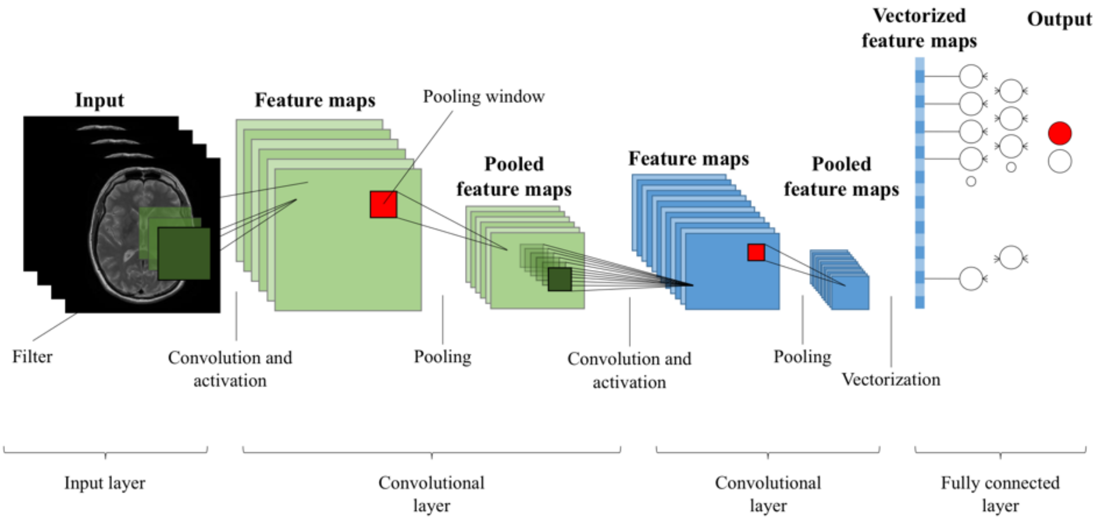


Fig. 1: The architecture of a convolutional neural network (CNN) model used in medical image classification. (Modified from the Figure in [1])

algorithm and demonstrated the use of the proposed framework on two public datasets: PPMI and OASIS;

- We demonstrated the impact of the data division strategy on the model performances by comparing the results based on two different split approach, one of which affected by data leakage.

This paper is organized as follows: In Section II, there is an overview on the related work. Section III describes the steps of the proposed methodology in detail. Classification results are presented in Section IV and discussed in Section V. Finally, Section VI concludes the paper with some remarks and indicates possible future directions.

II. Related Work

PD is a neurological disorder caused by the progressive death of dopamine producing cells in the brain [29], [30]. It is the second most common neurodegenerative disorder after Alzheimer's Disease (AD) [31]. An estimated 7 to 10 million people worldwide have been affected by PD and related disorders in 2018 [32].

In recent years, several neuroimaging studies have utilized machine learning (ML) algorithms for detection and diagnosis of PD [33]–[35]. Various modalities like Magnetic Resonance Imaging (MRI), Single Photon Emission Computed Tomography (SPECT), Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI) are used within these research to diagnose PD [36], [37]. In 2018, Esmaeilzadeh *et al.* [22] used 3D CNN for simultaneous classification and regression of PD diagnosis based on MRI and personal information (i.e., age and gender). They achieved 100% accuracy on both test and validation sets. In that study, they reached to the conclusion that Superior Parietal part on the right hemisphere of the brain is very critical in the diagnosis of PD. Lei *et al.* [38] performed a multi-class classification of three different clinical statuses: PD, SWEDD, and healthy conditions (HC) via SVM. They concluded that the classification performance with multi-modality features (GCD) combined with cerebrospinal fluid (CSF) biomarkers and clinical scores (DSSM) is always better than those without additional features. Recently, Sivaranjini *et al.* [21] utilized AlexNet to diagnose PD. The image dataset with 80% of the input data are used for training, and the remaining 20% is used for testing. Through TL, they achieved an accuracy of 88.9% on the classification of MRI slices. However, they did not test their model with subjects that were not included in the training data.

AD, on the other hand, is the most common neurodegenerative disorder [39]. It is predicted that by 2050, half (51%) of all people 65 and older will be facing with AD [40].

Sarraf *et al.* [26] used a CNN model for AD diagnosis in adults (above 75 years old) using fMRI and MRI. The data was divided into three parts: training (60%), validation (20%), and test (20%). They achieved 99.9% accuracy for functional MRI data and 98.84% for MRI data, respectively. However, data division was not done at the subject-level leading data from the same subject to be in both the training and test sets.

In [28], Payan and Montana designed a classification system that combines sparse autoencoders and convolutional neural networks. They divided ADNI dataset into training set (1,731 samples), validation set (306 samples) and test set (228 samples) and achieved 95.39% classification accuracy with both 2D CNNs and 3D CNNs. Again, they did not perform subject level division. Lastly, Hon *et al.* [24] utilized two state-of-the-art architectures, namely VGG16 and Inception V4 to classify AD. They used 5-fold cross-validation to obtain the results, with an 80% - 20% split between training and test. By using a pre-trained model for transfer learning (TL), they reported 92.3% accuracy with VGG16 model and 96.25% with Inception model.

When we check the literature, we see that the phenomenon

TABLE I: Summary of the studies with potential of data leakage. Studies perform Parkinson's Disease (PD) and Alzheimer's Disease (AD) classification using 2D or 3D convolutional neural networks (CNNs) with structural magnetic resonance imaging.

Disease	Study	No. of subjects	No. of MRIs	Data division method	Accuracy (%)
PD	Sivaranjini et al., 2019 [21]	182	7646 slices (2D)	4:1 train/test split by MRI slices	88.9
PD	Esmaeilzadeh et al. 2018 [22]	452	452 volumes (3D)	8.5:1:0.5 train/development/test split by augmented patches from MRI	100
AD	Jain et al.,2019 [23]	150	3000 slices (2D)	8:2 train/test split, by augmented MRI slices	95
AD	Hon and Khan, 2017 [24]	200	6400 slices (2D)	4:1 train/test split by MRI slices, 5-fold cross-validation	92.3
AD	Farooq et al., 2017 [25]	355	38024 slices (2D)	3:1 train/test split by MRI slices	98.8
AD	Sarraf and Toghi, 2016 [26]	n/a	90300 slices (2D)	3:1:1 train/validation/test split, 5-fold cross validation	96.85
AD	Wu et al., 2018 [27]	457	21936 slices (2D)	2:1 train/test split, 5-fold cross validation	97.58
AD	Payan and Montana, 2015 [28]	n/a	100 volumes (3D)	8:1:1 train/validation/test split, by patches from MRI	89.47

known as data leakage, is indeed a serious problem. Still, many papers published in the area are suffering from biased results most probably caused by limited experience with medical data. While working on this paper, we became aware of the recent work by Wen *et al.* [41] that illustrated the presence of data leakage across various studies which use ML in AD classification. They performed a rigorous literature search on AD and grouped the studies into three categorize: (a) studies without data leakage; (b) studies with potential data leakage and (c) studies with clear data leakage. They observed data leakage in 42% of surveyed papers.

III. Methods

In this section, we briefly describe the datasets we have used, the pre-processing steps and finally, the model architectures together with training protocols.

A. Data Splitting

Throughout the work, we realized that a common misconception occurs in many different papers which use machine learning algorithms in 3D medical imaging. Performance of the models was often determined by dividing the pooled slices into training and test sets [21], [24]–[26], [42] (see Table I). Thus, training and test sets included the different brain slices of the same subjects. Unfortunately, in that case, the high accuracies may stem from high intra-subject correlation. To test our hypothesis, we employed two different data splitting approaches. First, we divided the data by subject, in which all of the MRI slices of a subject are placed either in the training or in the test set. Then, in the second part, we pooled all slices together and then split the overall set randomly, meaning that the different slices of the same patient could appear both in the training and test sets.

B. Datasets

In this study two datasets were used, namely Parkinsons Progression Markers Initiative (PPMI) database [43] for PD and Open Access Series of Imaging Studies (OASIS) [44] for AD.

1) PPMI: The axial T2 weighted MRI slices used to classify PD in this work are from the PMMI database (Table II). The reason behind using T2 weighted MRI for PD is that T2 weighted sequences are better at detecting changes in tissue properties [45]. As a result, the data has the potential to monitor the structural changes of the brain caused by PD, such as the reduced volume of caudate and putamen [46].

The PPMI database is publicly available and helps researchers to conduct research on identifying biomarkers of PD progression. It consists of a set of three-dimensional brain slices of 452 PD patients (292 males and 160 females) and 204 HC (134 males and 70 females). The average age of the patients is 61, where the minimum age is 30, and the maximum age is 89.

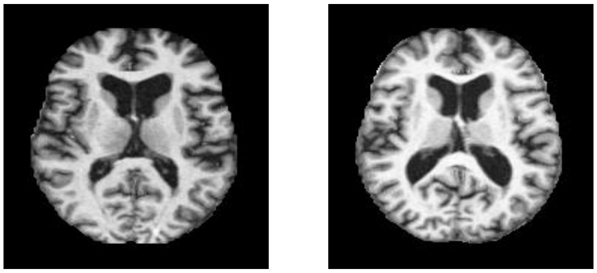
The PPMI subset used in this study consists of 408 subjects with 204 HC and 204 PD subjects. It has 6569 MRI slices derived from HC and 4467 slices from PD subjects. We randomly picked 7030 slices in total for our slice-based PD subset. Of these, 3515 slices were PD, and the remaining 3515 were HC. For the random division case, we used 80% of these slices in the training process while the rest were assigned to the test set. For the subject based case, we divide the data by patient meaning that the MRI slices of 164 patients from each class are placed in the training set and the slices of 40 AD patients and 40 HC are assigned to the test set.

2) OASIS: For classification of AD, we used cross-sectional, structural MRI data from the OASIS database (Table III). For the random split tests, we have employed the exact data set which were used in Hon *et al.*'s work [24] in order to replicate their approach while avoiding bias.[1] The subset they have used in their work consists of cross-sectional T1-weighted MRI scans. In their experiments, they randomly picked 200 subjects, 100 of whom were chosen from the AD group, while the other 100 from the HC group. The sample MRI slices from OASIS data can be seen in Figure 2.

For the subject based case, we created a similar subset from the OASIS database by picking 200 subjects, half of whom were AD patients, while the other half was HC. MRI slices of 80 subjects from each class are used to train the model, while the other subjects took part in testing process. MRI scans from OASIS database are in hdr/img file format. To pre-process the scannings, we first converted them into NIfTI format, then into 2D (jpg) format.

The decision criteria of AD is a variable called Clinical Dementia Rating (CRD) with 0 suggests HC and any value greater than 0 implies AD. OASIS-1 dataset includes two different data: Raw and processed. Processed images are the

[1] The subset Hon *et al.* created from the OASIS data are accessible at https://github.com/marciahon29/Ryerson_MRP



(a) A sample magnetic resonance imaging slice of a Alzheimer's disease patient

(b) A sample magnetic resonance imaging slice of a health control

Fig. 2: Example of two Magnetic resonance imaging (MRI) slices of an Alzheimer's Disease (AD) subject and healthy control (HC) from OASIS database.

TABLE II: Demographic information of PMMI dataset.

Classes	No. of subjects	Sex	Age	No. of MRI slices
PD	204	101 M, 103 F	30-89	3015
HC	204	134 M, 70 F	30-89	3015

TABLE III: Demographic information of OASIS-1 dataset.

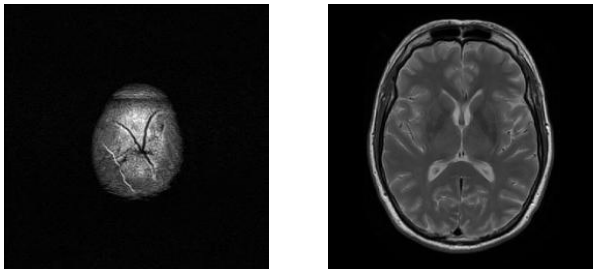
Classes	No. of subjects	Sex	Age	No. of MR slices
AD	100	65 M, 35 F	18-96	3200
HC	100	38 M, 62 F	18-96	3200

brain-masked version of atlas registered image that are used in both types of experiments.

C. Image Pre-processing

The input of the 2D CNNs that we utilized in our approach is the set of 2D slices extracted from the MRI volume. Typically, each MRI volume contains many slices that correspond to a different cross section of the brain. To increase the performance of classification, we decided to pick the most informative slices to train the network. It is known that a significant grey matter intensity loss with changes in the striatum region is observed in PD when compared with HC [46]. By calculating the image entropy for each slice, we aimed to select the slices which can illustrate such degenerated structure [24]. Two sets of MRI slices that belong to a PD patient are shown in the Figure 3. The slice on the left of the figure is not very informative in terms of the amount of gray matter it reveals when compared to the slice on the right.

Entropy is a measure of histogram dispersion which illustrates the variation in a slice. In the case of an image which has been perfectly histogram equalized, all 256 such states are equally occupied, and the entropy of the image is maximum. On the other hand, if all of the pixels of an image have the same value, the entropy is zero. Therefore, if the entropy of the image is reduced, its information is reduced as well. Thus, to obtain the most informative slices for network training, an entropy threshold has been determined (4.5, based on our empirical analysis).



(a) Non-informative slice in terms of the amount of the gray matter visible

(b) Informative slice in terms of the amount of the gray matter visible

Fig. 3: Example of two magnetic resonance imaging (MRI) slices of a Parkinson Disease (PD) subject.

For a slice, the entropy can be calculated as follows:

$$H = -\sum_{i=1}^{M} p_i \log p_i$$

where M is the number of gray levels (256 for 8-bit images) and p_i is the probability of a pixel having gray level intensity.

After eliminating the slices which fail to carry the necessary information, normalization is performed on the remaining MRI slices to obtain an unvaried contrast and intensity range. For this reason, each MRI slice in the data set is normalized to the range $(0, 1)$. To be compatible with the pre-trained models of VGG16 and Resnet50, the slices were resized to be 224×224.

We followed the same pre-processing structure for the AD slices as well.

D. CNN Models

We utilized two different architectures (VGG and ResNets) which are widely used in disease detection frameworks.

1) VGG16: VGG16 is a 16-layer network built by Oxfords Visual Geometry Group (VGG) and presented in their paper entitled "Very Deep Convolutional Networks for Large-Scale Image Recognition" [47]. It won the ImageNet competition in ILSVRC-2014 with the accuracy of 92.7%. It replaces large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) in the Alexnet with multiple 33 kernel-sized filters.

The input to the first layer is a fixed-size 224 × 224 RGB image. The image is then passed through a stack of convolutional layers as well as max pooling layers. Finally, convolutional layers are followed by three Fully-Connected (FC) layers and the soft-max layer for 1000-way ILSVRC classification. The architecture of VGG16 is shown in the Figure 4.

2) Resnet50: Residual neural network (ResNet) ranked first in the ILSVRC 2015 classification competition with top-5 error rate of 3.57%. He *et al.* [48] ease the training process of deep neural networks while making their model deeper than those used previously. They reformulate the layers as learning residual functions with reference to the layer inputs, rather than learning unreferenced functions. Residual neural

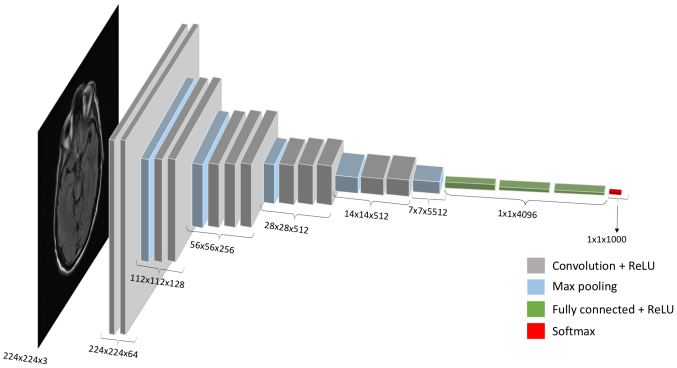


Fig. 4: The architecture of the VGG16 model adopted for magnetic resonance imaging (MRI) data.

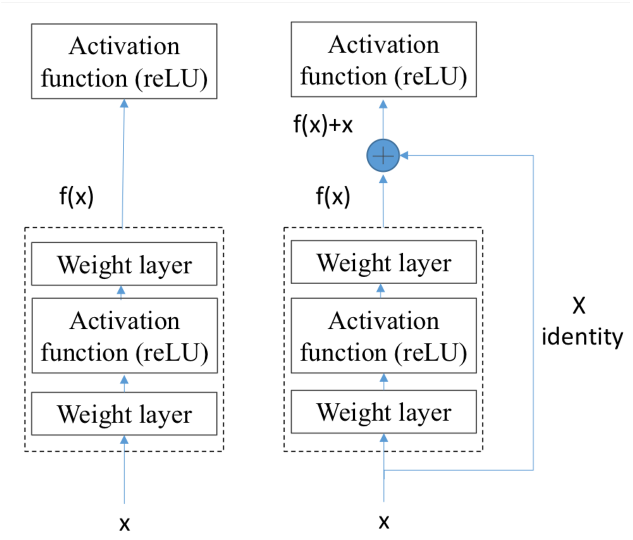

Activation function (reLU)
f(x)
Weight layer
Activation function (reLU)
Weight layer
x
Activation function (reLU)
f(x)+x
f(x)
Weight layer
Activation function (reLU)
Weight layer
X identity
x

Fig. 5: A building block of a regular learning (left) and a residual learning (right) (from He, 2016 [48]).

networks solve the problem known as vanishing gradient. When the network is too deep, the gradients of the loss function approaches zero, making the network hard to train. As a result, the weights are not updated, and thus learning cannot be achieved. With ResNets, the gradients can flow directly through the skip connections backward from latter layers to initial filters. The building block of a sample residual neural network structure is shown below in the Figure 5.

E. Training Protocols and Transfer Learning

Acquiring large sets of labeled data in medical imaging is a hard task as it is mostly sealed due to privacy and institutional policies, or expensive to label. To avoid the common problem of overfitting which generally stems from small data set and deep networks, transfer learning (TL) is employed to train a model efficiently on a smaller data set.

The idea behind TL is that many deep neural networks trained on images exhibit a common behavior: the first layers extract generic features and perform general operations such as edge detection or color blob detection [49]. Such low level features might be applicable to many datasets and tasks. Thus, when a network is pre-trained on an extremely large dataset, such as ImageNet, comprising 1.4 million images with 1000 classes, knowledge extracted from there can be applied to the given task of interest. Even for cross-domain application, such as networks trained on natural images used with medical images, TL has been proved to be robust [50].

TABLE IV: Tested architectures and their corresponding average accuracy on two dataset (PPMI and OASIS) using two data divisions (RD-Random Division, SbD-Subject-based Division).

	PD (Data 1: PPMI)		**AD (Data 2: OASIS)**	
	RD	SbD	RD	SbD
VGG16	82.8%	61.2%	90.47%	64.3%
Resnet50	88.6%	67.3%	92.5%	67.1%

For transfer learning, we follow the fine-tuning approach, where the last three layers of the pre-trained model are modified. The weights of the other layers of the model were frozen during fine-tuning to prevent overfitting. For VGG16, 50 epochs were used with a batch size of 40. The stochastic gradient descent and Adagrad optimization algorithms were used to minimize cross-entropy type of error. For Resnet50, 100 epochs with batch size of 32 were used. The optimization model was stochastic gradient descent. The loss function was categorical cross-entropy.

Data selection method and pre-processing part mentioned in Section III are implemented in MATLAB [51]. Then, deep learning methods are executed using Keras [52] with a TensorFlow [53] backend. Architectures as well as the pre-trained weights were available to download in open source repositories of the models.

IV. Experimental Results

The main aim was to differentiate AD and PD patients from HC by analyzing MRI data derived from two different databases via the CNN models and to show the importance of data division method on the generalization performance of the models. Table IV illustrates the accuracy results of the two models across two separate datasets using subject-level data splitting and random splitting after pooling all slices.

As it can be seen from the Table IV, both VGG16 and Resnet models can classify PD from HC with more than 82% accuracy when data are randomly split (biased split). However, on subject based split (unbiased split), we observed a large drop in accuracy (17% to 25%) for classification of the disease. Again, for AD classification, the same pattern can be detected. When data are divided at subject level, classification accuracy of VGG16 model is 64.3% whereas Resnet50 model achieves 67.1%. Alarmingly, pooling then splitting at slice level can inflate the classification accuracy by 26.1 percent points compared to the subject level split.

V. Discussion

Comparison of classification performances across studies is an arduous task as each study has various pre-processing stages, validation approach or hyperparameter selection. In studies which create subsets from publicly available datasets, the selection of the subset is often a random process, which makes it impossible to replicate the work accurately [24]. Moreover, some of the studies do not provide sufficient implementation details, especially about the validation procedures adopted, with the risk that the reported performances are affected by significant bias. Dividing the data at the slice-level in medical image classification is a significant problem which is currently widespread in the field. Our results show that this may artificially inflate the accuracy of classifiers by as much as 26 percentage points.

To evaluate prospective clinical feasibility of automated diagnosis, unbiased and accurate assessment of the model performances is crucial. We argue that despite the impressive accuracies of the previous works, there still exist some serious issues that must be resolved and much room for improvement in medical image classification and automated diagnosis.

VI. Conclusion

In this paper, we utilized a transfer learning-based method to detect two most common neurological diseases from structural MRI images. We employed two state-of-the-art architectures, namely VGG16 and Resnet, to classify PD subject from HC and AD subjects from HC. We test our models on MRI slices from the PMMI and OASIS brain imaging datasets, where MRI slices of more than 300 patients are used to train the models. We compared the results of two data split approaches across separate data sets, and showed that there is a large overestimation in accuracy when slices from all subjects are pooled together prior to validation.

The large discrepancy of accuracies between two types of data division suggests that the test accuracy from the random division approach is not a valid measure of performance on new subjects. Subject level tests are required to show the accurate performance of the classification model.

While we are confident that most researchers are well aware of the issue and would never split data from the same subject into test and training data, we have found that this is still a serious problem in the literature. With the recent advances in machine learning and AI, more and more people are becoming interested in applying these techniques to biomedical imaging and there is a real and growing risk that many of them will not be familiar with the possible issues and the good practices.

In the future, we will investigate other state of the art models as well as the effect of deep fine tuning on performance. Optimizing the hyperparameters of the models and expending the datasets via collaborations may be crucial to achieve better results. With these efforts, we aim to solve the problem behind the low accuracy of subject level tests. We hope to achieve better patient group classication and ease the diagnosis of neuro-degenerative disorders in the near future.

References

[1] R. C. Gonzalez and R. E. Woods, "Image processing," *Digital image processing*, vol. 2, p. 1, 2007.

[2] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," *Artificial intelligence in medicine*, 2018.

[3] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," *PeerJ*, vol. 7, p. e6201, 2019.

[4] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Computer methods and programs in biomedicine*, vol. 157, pp. 19–30, 2018.

[5] M. Kirienko, M. Sollini, G. Silvestri, S. Mognetti, E. Voulaz, L. Antunovic, A. Rossi, L. Antiga, and A. Chiti, "Convolutional neural networks promising in lung cancer t-parameter assessment on baseline fdg-pet/ct," *Contrast Media & Molecular Imaging*, vol. 2018, 2018.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[7] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[8] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[9] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do cifar-10 classifiers generalize to cifar-10?," *arXiv preprint arXiv:1806.00451*, 2018.

[10] A. Blum and M. Hardt, "The ladder: A reliable leaderboard for machine learning competitions," *arXiv preprint arXiv:1502.04585*, 2015.

[11] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, vol. 2, no. 3, pp. 249–262, 2018.

[12] J. Larsen and C. Goutte, "On optimal data split for generalization estimation and model selection," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98TH8468)*, pp. 225–234, IEEE, 1999.

[13] S. Lohr, "Sampling: Design and analysis.,(brooks/cole publishing: Pacific grove, ca.)," 1999.

[14] G. J. Bowden, H. R. Maier, and G. C. Dandy, "Optimal division of data for neural network models in water resources applications," *Water Resources Research*, vol. 38, no. 2, pp. 2–1, 2002.

[15] R. D. Snee, "Validation of regression models: methods and examples," *Technometrics*, vol. 19, no. 4, pp. 415–428, 1977.

[16] J. E. Trost, "Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies," *Qualitative sociology*, vol. 9, no. 1, pp. 54–57, 1986.

[17] N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nature neuroscience*, vol. 12, no. 5, p. 535, 2009.

[18] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, 2017.

[19] A. Torralba, A. A. Efros, *et al.*, "Unbiased look at dataset bias.," in *CVPR*, vol. 1, p. 7, Citeseer, 2011.

[20] A. Ashraf, S. Khan, N. Bhagwat, M. Chakravarty, and B. Taati, "Learning to unlearn: Building immunity to dataset bias in medical imaging studies," *arXiv preprint arXiv:1812.01716*, 2018.

[21] S. Sivaranjini and C. Sujatha, "Deep learning based diagnosis of parkinsons disease using convolutional neural network," *Multimedia Tools and Applications*, pp. 1–13, 2019.

[22] S. Esmaeilzadeh, Y. Yang, and E. Adeli, "End-to-end parkinson disease diagnosis using brain mr-images by 3d-cnn," *arXiv preprint arXiv:1806.05233*, 2018.

[23] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based alzheimers disease classification from magnetic resonance brain images," *Cognitive Systems Research*, vol. 57, pp. 147–159, 2019.

[24] M. Hon and N. M. Khan, "Towards alzheimer's disease classification through transfer learning," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1166–1169, IEEE, 2017.

[25] A. Farooq, S. Anwar, M. Awais, and S. Rehman, "A deep cnn based multi-class classification of alzheimer's disease using mri," in *2017 IEEE International Conference on Imaging systems and techniques (IST)*, pp. 1–6, IEEE, 2017.

[26] S. Sarraf and G. Tofighi, "Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks," *arXiv preprint arXiv:1603.08631*, 2016.

[27] C. Wu, S. Guo, Y. Hong, B. Xiao, Y. Wu, Q. Zhang, A. D. N. Initiative, *et al.*, "Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks," *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 10, p. 992, 2018.

[28] A. Payan and G. Montana, "Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks," *arXiv preprint arXiv:1502.02506*, 2015.

[29] R. Postuma and J. Montplaisir, "Predicting parkinson's disease–why, when, and how?," *Parkinsonism & related disorders*, vol. 15, pp. S105–S109, 2009.

[30] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor parkinson's disease severity symptoms using a machine learning approach," *Artificial intelligence in medicine*, vol. 58, no. 3, pp. 195–202, 2013.

[31] S. Halbgebauer, M. Nagl, H. Klafki, U. Haußmann, P. Steinacker, P. Oeckl, J. Kassubek, E. Pinkhardt, A. C. Ludolph, H. Soininen, *et al.*, "Modified serpina1 as risk marker for parkinsons disease dementia: Analysis of baseline data," *Scientific reports*, vol. 6, p. 26145, 2016.

[32] U. Parkinsons, "The incidence and prevalence of parkinsons in the UK," *London, UK*, 2018.

[33] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices," *Npj Digital Medicine*, vol. 1, no. 1, p. 39, 2018.

[34] G. An, K. Omodaka, K. Hashimoto, S. Tsuda, Y. Shiga, N. Takada, T. Kikawa, H. Yokota, M. Akiba, and T. Nakazawa, "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," *Journal of healthcare engineering*, vol. 2019, 2019.

[35] T. A. Shaikh and R. Ali, "Applying machine learning algorithms for early diagnosis and prediction of breast cancer risk," in *Proceedings of 2nd International Conference on Communication, Computing and Networking*, pp. 589–598, Springer, 2019.

[36] G. Garraux, C. Phillips, J. Schrouff, A. Kreisler, C. Lemaire, C. Degueldre, C. Delcour, R. Hustinx, A. Luxen, A. Destée, *et al.*, "Multiclass classification of fdg pet scans for the distinction between parkinson's disease and atypical parkinsonian syndromes," *NeuroImage: Clinical*, vol. 2, pp. 883–893, 2013.

[37] M. Tahmasian, L. M. Bettray, T. van Eimeren, A. Drzezga, L. Timmermann, C. R. Eickhoff, S. B. Eickhoff, and C. Eggers, "A systematic review on the applications of resting-state fmri in parkinson's disease: does dopamine replacement therapy play a role?," *Cortex*, vol. 73, pp. 80–105, 2015.

[38] H. Lei, Y. Zhao, Y. Wen, Q. Luo, Y. Cai, G. Liu, and B. Lei, "Sparse feature learning for multi-class parkinsons disease classification," *Technology and Health Care*, vol. 26, no. S1, pp. 193–203, 2018.

[39] S. Hague, S. Klaffke, and O. Bandmann, "Neurodegenerative disorders: Parkinsons disease and huntingtons disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 8, pp. 1058–1063, 2005.

[40] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[41] J. Wen, E. Thibeau-Sutre, J. Samper-Gonzalez, A. Routier, S. Bottani, S. Durrleman, N. Burgos, and O. Colliot, "Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation," *arXiv preprint arXiv:1904.07773*, 2019.

[42] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Classification using deep learning neural networks for brain tumors," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 68–71, 2018.

[43] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, *et al.*, "The parkinson progression marker initiative (ppmi)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[44] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[45] B. Heim, F. Krismer, R. De Marzi, and K. Seppi, "Magnetic resonance imaging for the diagnosis of parkinsons disease," *Journal of neural transmission*, vol. 124, no. 8, pp. 915–964, 2017.

[46] U. Saeed, J. Compagnone, R. I. Aviv, A. P. Strafella, S. E. Black, A. E. Lang, and M. Masellis, "Imaging biomarkers in parkinsons disease and parkinsonian syndromes: current and emerging concepts," *Translational neurodegeneration*, vol. 6, no. 1, p. 8, 2017.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[48] 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, 2016.

[49] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[50] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 1–6, 2018.

[51] MATLAB, *version 9.5.0.944444 (R2018b)*. Natick, Massachusetts: The MathWorks Inc., 2018.

[52] F. Chollet, "keras." https://github.com/fchollet/keras, 2015.

[53] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.