Proceedings of Machine Learning Research 183:163–177, 2022 Short Paper - LIDTA 2022

CNN and diffusion MRI's 4th degree rotational invariants for Alzheimer's disease identification

Aymene Mohammed Bouayed AYMENE.BOUAYED@INRIA.FR
Aix Marseille Université, Centre Inria d'Université Côte d'Azur, France

Samuel Deslauriers-Gauthier SAMUEL.DESLAURIERS-GAUTHIER@INRIA.FR
Mauro Zucchelli MAURO.ZUCCHELLI@INRIA.FR
Rachid Deriche RACHID.DERICHE@INRIA.FR
Centre Inria d'Université Côte d'Azur, France

For ADNI *

Editor: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak and Shuo Wang.

Abstract

Recently, a general analytical formula to extract all the Rotation Invariant Features (RIFs) of the diffusion Magnetic Resonance Imaging (dMRI) signal was proposed. The features extracted using this formula represent a generalisation of the usual second degree RIFs such as the mean diffusivity. In this work, we study the usefulness of all the 12 algebraically independent RIFs extracted from 4th degree spherical harmonics that model the dMRI signal per voxel in the context of Alzheimer Disease (AD) identification. To do so, and since we are working with imbalanced data sets, we first introduce a non-linear metric to evaluate the performance of the models, the (*B-score*). This proposed metric allows high score only when both classes are distinguished correctly. We use the proposed metric in conjunction with a deep Convolutional Neural Network that operates on subject slices to identify if a subject has AD or not. We find that micro-structure information communicated by RIFs is indeed useful to AD identification and that not all RIFs are equivalently useful. We also identify the two best RIF combinations for the ADNI - SIEMENS and the ADNI - GE medical data sets respectively. The combination of these RIFs achieves a classification B-score of 73.62% and 72.31% on the previous data sets respectively. We note the importance of combining high degree RIFs with low degree ones to improve the classification performance.

Keywords: Convolutional neural network, imbalanced data set evaluation metric, dMRI, Rotation invariant features.

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative(ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. Introduction

Alzheimer's disease (AD) is one of the incurable neuro-degenerative diseases. It consists of dying neurons in the brain which manifests as dementia and it is observed on Magnetic Resonance Images (MRI) as a shrinking of the size of the white and gray matter and an increase in the size of the Cortico-Spinal Fluid (CSF) of the ventricles (Jones, 2012). It is difficult to get an early diagnosis of this illness via only visual interpretation of MRI. As a consequence, many deep learning approaches have been proposed such as those of Khan et al. (2019); Chitradevi and Prabha (2020); Ding et al. (2019); Suh et al. (2020); Loddo et al. (2022). All of these methods use MRI scans to identify AD as it is the only clinically used mean. Thus, when working with medical data the problem of imbalanced data sets always arises. This is because the prevalence of diseases is low compared to healthy patients. To resolve this problem, previous works rely on data augmentation in order to increase the variability and the number of data points in each class so as to have the same number of data points or manually make sure the data sets are balanced. Then, they fine-tune on slices of MRI scans a deep learning model such as Visual Geometry Group (VGG) model Simonyan and Zisserman (2014), Inception V3 model Szegedy et al. (2015), AlexNet Krizhevsky et al. (2012), Resnet-101 He et al. (2015) or Inception Resnet V2 Szegedy et al. (2016). Khan et al. (2019) rely on intelligent slice selection to pick-out the most informative slices. Other methods such as Loddo et al. (2022) use ensemble learning methods to improve the classification performance. When using MRI data the previous works achieve very high classification performances ($\geq$ 90% accuracy). This is because MRI scans reflect the anatomical structure of the brain and models can see the shrinking of the brain's white and gray matter as well as the increase in the CSF volume.

In this work, we investigate the brain's white matter changes associated with AD. To this end, the use diffusion MRI (dMRI) data while avoiding anatomical clues by registering all subject scans to the same space. Few studies have investigated the use of dMRI to tackle the task of AD identification, with the exception of the work of Aderghal et al. (2018). The use of dMRI data instead of anatomical MRI is of interest because it provides information on the brain's micro-structure which can be used to infer the brain's white and gray matter changes (Jones, 2012).

In Aderghal et al. (2018) the authors propose to pre-train a neural network to classify on slices of structural Magnetic Resonance Imaging (sMRI) data. Then, fine-tune the model's weights on slices of Mean Diffusivity (MD). The MD slices correspond to the average water molecules diffusivity per voxel and are recovered from dMRI data using the DTI acquisition scheme (Jones, 2012). Thus, the authors only use a 28×28 patch that contains the hippocampal region and only consider the MD which may not be the optimal measure of changes in the brain's micro-structure. The reported classification accuracy obtained using this method on the ADNI data set is 92.5%. Furthermore, the work of Aderghal et al. (2018) uses as input sMRI and MD slices that are registered using linear registration to the Montreal Neurological Institute space. This type of registration deals with translation and rotation transformations which preserve the anatomical shape of the brain. Consequently the anatomical clues of AD are still present in the data.

In this work, we rely on all the 12 algebraically independent Rotation Invariant Features (RIFs) extracted from 4th degree Spherical Harmonics that model the dMRI signal

(Zucchelli et al., 2020) to quantify white matter changes associated AD. RIFs capture the form of the diffusion signal per voxel while avoiding its different rotations and they have been shown to be linked to the local brain micro-structure in Zucchelli et al. (2020). Also, RIFs can be seen a form of dimensionality reduction on the space of the acquired diffusion signal. This is of particular interest in our work where we do not have a lot of subjects.

Indeed, in this work we answer the following 4 questions :

1. How can we evaluate a model on an imbalanced data set while enforcing greater balance between the classes?
2. Is the RIFs' micro-structure information helpful for AD identification?
3. Are there RIFs that are more important than others?
4. Does the combination of the best RIFs improve the classification performance?

Firstly, to answer first question, we propose a metric that is fair in the context of imbalanced data sets and helps to fairly measure the models' performances. Since most medical imaging data sets are imbalanced in favor of healthy subjects because the prevalence of the majority of diseases is low, the use of the accuracy metric does not provide much information on the performance of the models where if a model classifies every data point in the majority class the accuracy will be high even though the model did not learn any decision boundary. Previous works report the true positive and the true negative rates, thus it is difficult to keep track of the balance of both metrics. The balanced accuracy metric regroups both rates in a simple linear formula. Thus due to it's linearity it does not perform well and the scores it provides can be interpreted in different ways. Consequently, we propose the *B-score*, a non-linear metric that combines the true and false positive rates and can be generalised to a multi class setting using the fractions of correctly classified data points w.r.t. the number of data points in that class. Secondly, To answer the second and third questions, we use each one of the RIFs as input to the proposed deep neural network to do subject classification into AD or NC. Finally, we experiment with combining the best RIFs of each data set to answer question four.

Subsequently, we summarise our contributions in this work as follow :

- The *B-score* : a fair and intuitive metric to evaluate models on imbalanced data sets.
- A deep learning model that operated on slices of RIFs to circumvents the problem of over-fitting and achieves a classification performance of 73.62% B-score on the ADNI - SIEMENS data set using the RIF combination $R_{22} \bigoplus R_{2444}$ and 72.31% B-score on the ADNI - GE medical data set using the RIF combination $R_{44} \bigoplus R_{2224} \bigoplus R_{2244}$.
- A study of the importance of each RIF and the combination of the best RIFs for the task of AD identification on multiple scanners.

2. Background/Theory

2.1. Diffusion MRI signal as a spherical function

Consider a set of dMRI signals acquired in a voxel using multiple gradient directions at the same b-value (HARDI acquisition scheme (Descoteaux, 2008)). The signal can be ap-

proximated using spherical harmonics (SH) of maximum degree D. The signal can then be interpolated in any gradient direction using:

$$f(\theta,\phi) = \sum_{l=0,\ \text{even}}^{D} \sum_{m=-l}^{l} c_l^m Y_l^m(\theta,\phi) \tag{1}$$

where θ and ϕ are the azimuth and inclination angles respectively. These angles represent a position on a unitary sphere in spherical coordinates. $Y_l^m(\cdot)$ are the real SH given by

$$Y_l^m = \begin{cases} \sqrt{2}\cdot Re(y_l^{|m|}), & \text{if } m > 0 \\ Re(y_l^0), & \text{if } m = 0 \\ \sqrt{2}\cdot Img(y_l^{|m|}) & \text{if } m < 0 \end{cases}$$

where

$$y_l^m(\theta,\phi) = i^{m+|m|}\sqrt{\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}} P_l^{|m|}(\cos\theta).\exp(im\phi).$$

c_l^m are the scalar coefficients of SH that are estimated from the acquired data, l and m represent the degree and order of the spherical function respectively. Only real SH are used because the diffusion signal is real so their is no need to use the complex form of the SH basis. Also, since the diffusion signal is antipodaly symmetric, only even degree SH are considered.

2.2. Rotation Invariant Features

Using Equation (1) multiple Rotation Invariant Features (RIFs) can be extracted. A simple example of RIFs is the mean of the values on the sphere. The formula

$$R_{l_1,\cdots,l_d} = \sum_{m_1=-l_1}^{l_1} \cdots \sum_{m_d=-l_d}^{l_d} c_{l_1 m_1}\cdots c_{l_d m_d} G(l_1,m_1|\cdots|l_d,m_d) \tag{2}$$

$$\forall l_i \in \mathbb{N}\ s.t.\ l \bmod 2 \equiv 0.$$

developed by Zucchelli et al. (2020) can be used to find the set of all RIFs up to a maximum degree d, where $G(\cdot)$ represent the Gaunt coefficients, $c_{l_i m_i}$ are the coefficient from Equation (1) linked to the degree l_i and order m_i. In the same work, Zucchelli et al. (2020) identified the set of all algebraically independent RIFs up to degree $d = 4$ which amounts to 12 RIFs and they show that these RIFs are directly linked to the underlying local brain micro-structure.

3. Data and Method

This works adopt a deep learning approach where a Convolutional Neural Network (CNN) is trained on the extracted and normalised RIFs to identify if a patient has the Alzheimer Disease (AD) or has a Normal brain Connectivity (NC). To do so, we start from raw dMRI data that needs to be preprocessed beforehand including the RIF extraction and slicing steps. Finally, we train a CNN model to classify patients into AD or NC. An overview of the proposed pipeline is illustrated in Figure 4 of Appendix D.

3.1. Data set

The data is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), diffusion magnetic resonance imaging (dMRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The data was acquired with different scanners such as GE medical and SIEMENS. Since the RIFs are extracted from dMRI data, we use the corresponding data sets and make the distinction between scans acquired using the SIEMENS scanner and the ones acquired using the GE medical scanner. This is because different manufacturers have different specifications. The number of available scans is reported in Table 1.

Table 1: Number of scans in the ADNI dataset by scanner type.

Scanner	Before preprocessing		After preprocessing	
	#AD scans	#NC scans	#AD scans	#NC scans
SIEMENS	51	378	46	352
GE medical	194	427	191	419

For the SIEMENS scanner the scans are given in the form of 55 volumes per subject, seven of them were acquired at a b-value of 0 and the rest at b-value 1000. As for the GE medical scanner, we have 46 volumes per subject, five of them were acquired at a b-value of 0 and the rest at b-value 1000. All the scans have been preprocessed and sliced according to the steps described in Appendix A and Appendix B.

3.2. B-score

From Table 1 it can be seen that, regardless of the scanner, we have an imbalanced data set in favor of the NC class. As a result, in order to evaluate the proposed model, we propose to use the *B-score* metric that we define as

$$\text{B-score} = n \times \frac{\prod_{i=0}^{n} frac\_class(i)}{\sum_{i=0}^{n} frac\_class(i)}. \tag{3}$$

where n is the number of classes and $frac\_class(i)$ is the number of subjects of class i that are correctly classified divided by the number of subjects of class i. In a two class classification setting, as is the case in this work, Equation (3) can be expressed as follow :

$$\text{B-score} = 2 \times \frac{frac\_class(AD) \times frac\_class(NC)}{frac\_class(AD) + frac\_class(NC)} = 2 \times \frac{TPR \times TNR}{TPR + TNR}$$

where TPR and TNR represent the True Positive Rate (sensitivity) and the True Negative Rate (specificity) respectively.

Combining these the true positive rate and the true negative rate measures in the B-score metric quantifies the classification of each class without neglecting the others. Indeed, this metric penalises the model more than the accuracy and balanced accuracy metrics when the classification is very poor in one class. As a consequence it enforces the balance of classification correctness for all classes compared to other metrics (See Figure 1). This can be illustrated in the extreme case using the data sets employed in this work and a model that classifies all AD and NC patients in the NC class. In this case, we would have $TPR = 0$ and $TNR = 1$ resulting in a balanced accuracy of 50%, the accuracy will be 68% and 88% on the test set of ADNI - SIEMENS and ADNI - GE medical data sets respectively, thus the B-score would be 0%. One advantage of the B-score over the F1-score is that it can be generalised to a multi-class setting. Another is that the F1-score is not symmetric and gives different results depending on which class is set as the positive one. This can also be observed in the setting were all data points are classified in one class where we would have an F1-score of 0% in one setting and > 0 in the other.

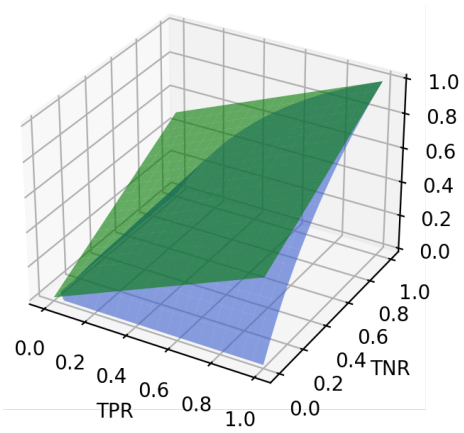


Figure 1: B-score's surface in blue vs Balanced accuracy's surface in green.

3.3. Deep learning model

The deep learning model we propose first uses convolutional layers to extract important information from the slices of a subject in input then aggregates these information into one vector that is used to take the final decision on the class of the subject.

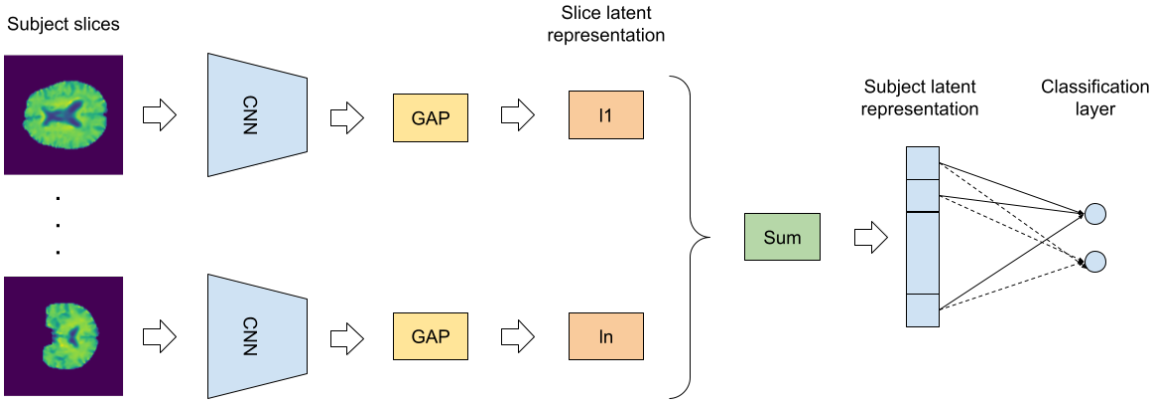


Figure 2: The proposed deep learning model.

More formally, let $\{x_i\}_{i=0}^{n}$ be a set of n slices of a subject X obtained using the Fixed slicing policy and f a Convolutional Neural Network (LeCun et al., 1998) with only convolutional layers. Using f we generate for each slice x_i a set of m features maps $F_i = \{F_{i,j}\}_{j=0}^{m}$. Then, we take the average of each feature map $F_{i,j}$ over j to get a latent representation

$l_i \in \mathbb{R}^m$ of the slice. This operation is called Global Average Pooling (GAP) (Lin et al., 2013). Finally, to get a latent representation l of the subject X we take the sum over all l_i (see Figure 2). This subject's latent representation l is fed to a Softmax classification layer to identify if the subject is of the AD or NC class.

Adopting this approach of global average pooling forces the network f to encode elementary input features in each feature map so that when we take the mean it represents the presence or absence of that particular feature. Also, by operating on slices we reduce the number of the parameters of the model which is a regularisation on the space of solution thus reducing the risk of over-fitting.

Intuitively, this approach can be seen as : if we have a feature that is present in any of the subject slices we encode it in the latent representation l. Then, this latent representation l is fed to a classification layer with two neurons to determine the class of the subject.

The architecture of the network f used in this work is composed of two convolutional blocks each one is composed of a 2D convolution with filters of size 4×4 and a stride of 2, batch normalisation, ReLU activation function then a dropout layer with a 20% probability. The first and second convolutional layers are composed of 50 and 75 filters respectively.

4. Results

In this section we present the different results of the used classification approach in different training settings using the B-score. We relegate the implementation details to Appendix C.

4.1. Single RIF performance

When looking at the difference in distribution of the RIFs across the classes as depicted in Figure 5 of Appendix D, we can see that their distributions can be differentiated as the distribution of the AD scans is shifted more towards null values which gives darker images.

To classify the subjects' scans, we rely on the deep learning model detailed in Section 3.3 to learn a decision boundary that best classifies the data as AD or NC based on a single RIF. Subsequently, we train the proposed model on each RIF independently on both the ADNI - SIEMENS and the ADNI - GE medical data sets for 50 epochs. We report the B-score obtained on the test set of each data set in Table 2. We note that all models converged or have been saved before over-fitting thanks to the check-pointing validation strategy. We do not include convergence curves due to space constraints.

Table 2: B-score obtained by each RIF extracted from the ADNI - SIEMENS and ADNI - GE medical data sets. The four best performing RIFs are highlighted in bold.

	R_0	R_{22224}	R_{2224}	R_{222}	R_{2244}	R_{224}
ADNI - SIEMENS	40.95%	**51.28%**	42.49%	**58.7%**	29.18%	0%
ADNI - GE medical	67.69%	64.40%	**72.29%**	66.61%	**69.82%**	66.24%
	R_{22}	R_{2444}	R_{244}	R_{4444}	R_{444}	R_{44}
ADNI - SIEMENS	**61.22%**	**60%**	0%	31.46%	40.63%	15.73%
ADNI - GE medical	67.60%	31.47%	**68.26%**	66.72%	50.56%	**74.28%**

From Table 2 it can be seen that some RIFs are more useful than others to the classification task at hand. For example, in the context of the ADNI - SIEMENS data set, the

RIF R_{22} achieves a test B-score of 61.22% whereas the RIF R_{224} achieves a B-score of 0% as it classifies every AD scan in the NC class. This emphasises the importance of using of the B-score metric where it quantifies more clearly the quality of the learned decision boundary.

Also, we observe a significant leap in performance for most RIFs when going from the ADNI - SIEMENS to the ADNI - GE medical data set. For example, R_{44} goes from 15% B-score to over 70%. Furthermore, we can see that the best RIFs for classification are not the same across the two tested data sets. Many factors could have this effect notably the small size and the scarcity of AD scans compared to NC scans in the ADNI - SIEMENS data set, noise levels in this data set, the fact that different manufacturers have different hardware specification and acquisition schemes. The study of these factors is relegated to a future work.

4.2. Multi-RIF performance

In this section, we train the proposed deep learning model for 150 epochs using 4 combinations of the previously best found RIFs of each data set namely R_{22}, R_{2444}, R_{222}, R_{22224} for the ADNI - SIEMENS data set and the RIFs R_{44}, R_{2224}, R_{2244}, R_{244} for the ADNI - GE medical data set. We start with only the best RIF then add in the second best as another channel in the input slice image, etc, ... The obtained test B-scores on both tested data sets are reported in Table 3.

Table 3: B-score obtained on the ADNI - SIEMENS and ADNI - GE medical data sets by each RIF combination denoted by the $\bigoplus$ symbol. The best performing RIFs combination on each data set is highlighted in bold.

(a) ADNI - SIEMENS data set.

RIF combination	B-score
R_{22}	72.99%
$R_{22} \bigoplus R_{2444}$	**73.62%**
$R_{22} \bigoplus R_{2444} \bigoplus R_{222}$	71.79%
$R_{22} \bigoplus R_{2444} \bigoplus R_{222} \bigoplus R_{22224}$	71.79%

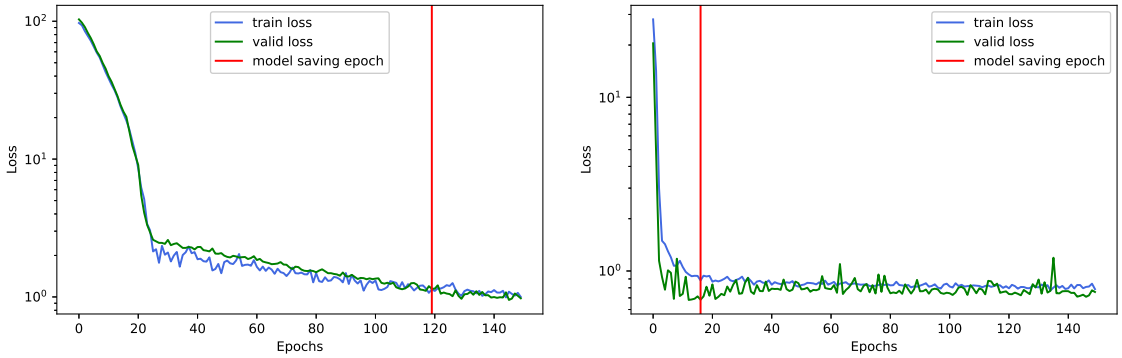
(b) ADNI - GE medical data set.

RIF combination	B-score
R_{44}	69.59%
$R_{44} \bigoplus R_{2224}$	66.24%
$R_{44} \bigoplus R_{2224} \bigoplus R_{2244}$	**72.31%**
$R_{44} \bigoplus R_{2224} \bigoplus R_{2244} \bigoplus R_{244}$	63.33%

From Table 3 we can identify that there is indeed a specific combination of RIFs that maximises the performance on each data set. For example, for the ADNI - SIEMENS data set we get the best test B-score of 73.61% when using the RIFs $R_{22} \bigoplus R_{2444}$ as for the ADNI - GE medical data set we get the best test B-score of 72.31% when using the RIFs $R_{44} \bigoplus R_{2224} \bigoplus R_{2244}$. Furthermore, it can be noticed that high order RIFs help improve the classification B-score. On both tested data sets we see that the best model combines a standard second order RIF be it R_{22} or R_{44} with a high order one(s). Thus, the RIFs combination that maximises the classification B-score of AD vs NC is not the same across the data sets.

In addition, when comparing Table 2 with Table 3, we see that when we train the proposed deep learning model for 50 epochs on the RIF R_{44} of the ADNI - GE medical data set we get a classification B-score of 74.28% whereas when we train for 150 epochs we get 72.31% B-score. This is due to the adopted model saving strategy where we save a model

weights that improve the classification performance on the validation set but not necessarily on the test set since we do not have access to it during the training phase and the chosen validation set's distribution in not representative enough of the test set's distribution.



(a) ADNI - SIEMENS data set. (b) ADNI - GE medical data set.

Figure 3: The log-scale evolution of the weighted cross entropy loss for train and validation of the proposed deep learning model trained on R22 $\bigoplus$ R2444 for the ADNI - SIEMENS data set and on R44 $\bigoplus$ R2224 $\bigoplus$ R2244 for the ADNI - GE medical data set.

Figure 3 represents the evolution of the training and validation loss of the best models from Table 3. We can see that these models did not over-fit to the training data sets and the best weights were saved at the lowest validation B-score point as enforced by the adopted saving protocol (See Appendix C). From these loss curves we can appreciate that this saving protocol allowed in the case of the ADNI - GE medical data set to store the weights just before the validation loss started to increase. As for the ADNI - SIEMENS data set, we can see that the weights of the model were saved just before the validation loss started to stabilise which backs the adopted model saving protocol.

5. Discussion and Conclusion

In this work, we studied the effectiveness of the 12 algebraically independent RIFs extracted from 4th degree spherical harmonics that model the dMRI signal in each voxel in the context of AD identification which is, to the best of our knowledge, the first work in this direction. The RIFs can be seen as a dimentionality reduction method on the space of the diffusion signal that extracts micro-structure information and helps cope with the lack of data. The closest work to ours is Aderghal et al. (2018) where the authors use dMRI data as MD slices which correspond to the RIF R_0. Thus, in Aderghal et al. (2018) the authors do not use the full brain scan, with only a low degree RIF: the MD. Also, they rely on the anatomical clues in the scan since they do not use non-linear registration. Whereas in our work, we try to gain a deeper understanding of the newly introduced RIFs general formula and high order RIFs Zucchelli et al. (2020) through data that is non-linearly registered to the same space to avoid anatomical clues and a simple CNN based pipeline that is trained using the weighted cross-entropy loss to take into consideration the fact the we are working with imbalanced

data sets. Also, we propose a novel metric to accurately asses the classification performance and reduce the impact of the imbalanced data set (Q1 of the introduction). Furthermore, we observe that RIFs contain micro-structure information that can be harnessed in the context of AD identification (Q2 of the introduction). Also, not all RIFs are important to the task of AD identification and there exists a combination of RIFs that maximises the classification performance. In addition, high order RIFs are very useful and contain information that is important to maximise the classification performance (Q3 and Q4 of the introduction).

A future extension to this work may include increasing the size of the data sets via Variational Auto-Encoder with a learned Riemannian latent space manifold as in Chadebec et al. (2021). Additionally, since in this work we avoided using anatomical clues in order to quantify the informativeness of the RIF, a future work could combine both approaches to maximize the classification performance which we can see that there is still room for improvement when using the B-score metric.

In conclusion, this work proves the usefulness of micro-structure information communicated by RIFs in the context of AD identification especially the high degree ones. Also, we propose a metric that quantifies the learned decision boundary more fairly and accurately.

This work can be found as a pre-print in the following link : https://hal.archives-ouvertes.fr/hal-03740103

Acknowledgments

This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Universite Cote d'Azur for providing resources and support.

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). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Karim Aderghal, Alexander Khvostikov, Andrei Krylov, Jenny Benois-Pineau, Karim Afdel, and Gwenaelle Catheline. Classification of alzheimer disease on imaging modalities with deep cnns using cross-modal transfer learning. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 345–350, 2018. doi: 10.1109/CBMS.2018.00067.

Clément Chadebec, Elina Thibeau-Sutre, Ninon Burgos, and Stéphanie Allassonnière. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder, 2021.

D. Chitradevi and S. Prabha. Analysis of brain sub regions using optimization techniques and deep learning method in alzheimer disease. *Applied Soft Computing*, 86:105857, 2020. doi: 10.1016/j.asoc.2019.105857.

Maxime Descoteaux. *High Angular Resolution Diffusion MRI: from Local Estimation to Segmentation and Tractography*. Theses, Université Nice Sophia Antipolis, February 2008.

Yiming Ding, Jae Ho Sohn, Michael G. Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W. Jenkins, Dmytro Lituiev, Timothy P. Copeland, Mariam S. Aboian, Carina Mari Aparici, Spencer C. Behr, Robert R. Flavell, Shih-Ying Huang, Kelly A. Zalocusky, Lorenzo Nardo, Youngho Seo, Randall A. Hawkins, Miguel Hernandez Pampaloni, Dexter Hadley, and Benjamin L. Franc. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290(2):456–464, 2019. doi: 10.1148/radiol.2018180958.

Rutger H.J. Fick, Demian Wassermann, and Rachid Deriche. The Dmipy Toolbox: Diffusion MRI Multi-Compartment Modeling and Microstructure Recovery Made Easy. *Frontiers in Neuroinformatics*, 13, October 2019. doi: 10.3389/fninf.2019.00064.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.09.015.

Derek K. Jones. *Diffusion MRI Theory, Methods, and Applications: Theory, Methods, and Applications*. Oxford University Press, Oxford, UK, 09 2012. ISBN 9780199965144. doi: 10.1093/med/9780195369779.001.0001.

Naimul Mefraz Khan, Marcia Hon, and Nabila Abraham. Transfer learning with intelligent training data selection for prediction of alzheimer's disease, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2013.

Andrea Loddo, Sara Buttau, and Cecilia Di Ruberto. Deep learning based pipelines for alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method. *Computers in Biology and Medicine*, 141:105032, 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.105032.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

C.H. Suh, W.H. Shim, S.J. Kim, J.H. Roh, J.-H. Lee, M.-J. Kim, S. Park, W. Jung, J. Sung, G.-H. Jahng, and for the Alzheimer's Disease Neuroimaging Initiative. Development and validation of a deep learning–based automatic brain segmentation and classification algorithm for alzheimer disease using 3d t1-weighted volumetric images. *American Journal of Neuroradiology*, 2020. ISSN 0195-6108. doi: 10.3174/ajnr.A6848.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

J-Donald Tournier, Robert Smith, David Raffelt, Rami Tabbara, Thijs Dhollander, Maximilian Pietsch, Daan Christiaens, Ben Jeurissen, Chun-Hung Yeh, and Alan Connelly. Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*, 202:116137, 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116137.

Mauro Zucchelli, Samuel Deslauriers-Gauthier, and Rachid Deriche. A computational framework for generating rotation invariant features and its application in diffusion mri. *Medical Image Analysis*, 60:101597, 2020. ISSN 1361-8415. doi: 10.1016/j.media.2019.101597.

Appendix A. Prepossessing

Since we are starting from raw dMRI data, a preprocessing phase is mandatory. In this phase, we follow the following steps :

1. ***Denoising*** to remove of some of the Racian noise present inherently in all dMRI data (Done using the MRtrix library (Tournier et al., 2019)).

2. ***Distortion correction*** to correct of distortions in the scans that may emerge from patients' movements (Done using the MRtrix library (Tournier et al., 2019)).

3. ***Rotation Invariant Features extraction*** to extract in each voxel all 12 algebraically independent RIFs calculated from 4th degree Spherical Harmonics that model the dMRI signal (Done using the dmipy library (Fick et al., 2019) which contrains the code from Zucchelli et al. (2020)).

4. ***Registration*** to remove anatomical clues and focus on the RIFs values in each voxel. It consists of casting all the scans onto one reference RIFs scan of a NC patient using a linear and non-linear transformations (Done using the FSL library (Jenkinson et al., 2012)).

After these steps, for each scan we obtain 12 corresponding 3D matrices each one represents one of the 12 RIFs. Note that not all scans passed the preprocessing step and the number of usable scans is reported in Table 1.

Also, we remove the outliers and normalise the values of each RIF 3D matrix of each subject. This is done my clipping the values in areas that constitute the brain into a suitable range specific to each RIF. Then, the values are normalised in the [0-1] range (Values outside the brain are set to 0).

Appendix B. Fixed Slicing

Due to the scarcity of dMRI scans compared to classical computer vision benchmark data sets, we cannot use each 4D matrix as a single data point in training a deep learning model that uses 3D convolutions as it will have too many parameters to learn. To solve this problem we resort to slicing the 4D RIF scans into 3D ones along the spatial directions, apply 2D convolutions on them, then regroup the resulting latent space (See Section 3.3). To this end, we take slices in the 2 main directions (Axial, Coronal) which results in a 3D matrix where the RIFs represent the channels [1]. As a slicing policy we opt for the *fixed slicing*. This policy consists of taking the same set of continuous slices in a predefined range regardless of the training epoch. We take all the brain slices that contain a reasonable amount of information then let the model decide which information to take/encode.

1. We have tried using the sagittal slices but they didn't improve the classification performance and nor allowed us to employ a big enough batch size.

Appendix C. Implementation details

C.1. Cross data set parameters and data set split

Building and training the deep learning model proposed in this work is done using the Pytorch library (Paszke et al., 2019). As for the optimisation part, it is done using the Adam optimisation algorithm (Kingma and Ba, 2014) with a learning rate of 2×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 10^{-5}. To split the data set into train, validation and test, we noticed that in the ADNI - SIEMENS data set the AD scans represent about 12% of the data set and in the ADNI - GE medical they represent around 32%. Consequently, we preserve this ratios in the train, validation and test splits of each data set. The data set partitioning we use is reported in Table 4.

Following the splits in Table 4, we have 62.5%, 13.5% and 24% of the data sets that is used for training, validation and testing respectively.

Table 4: ADNI data sets splits.

	ADNI - SIEMENS				ADNI - GE medical			
	Train	Validation	Test	Non used	Train	Validation	Test	Non used
AD	30	6	10	0	119	26	46	0
NC	200	46	76	30	262	54	100	3

C.2. Scans' size adjustments

For each data set we make sure that the voxels are isotropic. For the ADNI - GE medical data set we omit the first 50 and last 56 dimensions of the scans on the second and third axis because they do not contain any brain information. Then, we pad the volumes with zeros so as to have cubes (except for the dimension of the RIFs). Finally, we interpolate the output of the padding operation to the size $12 \times 224 \times 224 \times 224$.

C.3. Loss function

Since the training, validation and testing sets are imbalanced, to train the deep learning models we use the Weighted Cross Entropy loss (WCE) expressed as

$$\mathcal{L}_{\text{WCE}} = \sum_{i,j} \beta_i \cdot y_{i,j} \cdot \log(\hat{y}_{i,j})$$

where β_i is a scaler that represents the weight of the data point i, y_i is a one hot vector encoding the ground truth label and $\hat{y}_i$ is a vector representing the probability distribution predicted by the network for the input i over all the classes. The weight β_i is a hyper parameter that is set according to the proportion of the class y_i in the training set w.r.t the other classes so that all classes have the same number of training instances when scaled by β_i. In our case, we set it as

$$\beta_i = \begin{cases} 1 & \text{if } i \text{ is NC} \\ \frac{200}{30} \approx 6.67 & \text{if } i \text{ is of the ADNI - SIEMENS data set and is AD} \\ \frac{262}{119} \approx 2.2 & \text{if } i \text{ is of the ADNI - GE medical data set and is AD} \end{cases}$$

To test the trained models, we adopt a check-pointing strategy. In this strategy, and after 5 warm up epoch, we save the model's weights each time we improve the validation B-score. Then, for the testing phase, we load the latest saved model and evaluate it on the test set. Consequently, in case of over-fitting, this approach allows us to use the weights of the best learned model according to the validation set.

Appendix D. Additional Figures

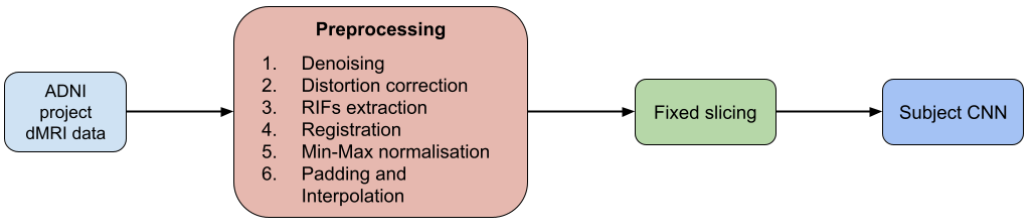
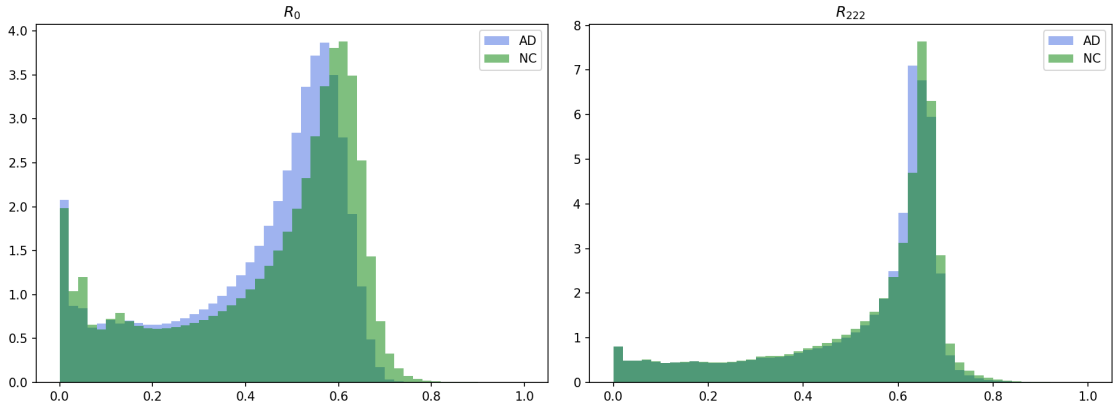


Figure 4: An overview of the proposed pipeline.



(*a*) AD vs NC for R_0.

(*b*) AD vs NC for R_{222}.

Figure 5: Distribution difference for the RIFs R_0 and R_{222} between the first 20 AD patients and the first 20 NC patient in the ADNI - SIEMENS data set.