NeuroImage: Clinical 14 (2017) 506–517

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

3D scattering transforms for disease classification in neuroimaging

Tameem Adel[a,*], Taco Cohen[a,b], Matthan Caan[c], Max Welling[a,b], On behalf of the AGEhIV study group and the Alzheimer's Disease Neuroimaging Initiative

[a]Machine Learning Lab, University of Amsterdam, The Netherlands
[b]Scyfer B. V., Amsterdam, The Netherlands
[c]Department of Radiology, Academic Medical Center (AMC), University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 12 August 2016
Received in revised form 29 January 2017
Accepted 3 February 2017
Available online 10 February 2017

Keywords:
Scattering representation
Feature extraction
MRI classification

ABSTRACT

Classifying neurodegenerative brain diseases in MRI aims at correctly assigning discrete labels to MRI scans. Such labels usually refer to a diagnostic decision a learner infers based on what it has learned from a training sample of MRI scans. Classification from MRI voxels separately typically does not provide independent evidence towards or against a class; the information relevant for classification is only present in the form of complicated multivariate patterns (or "features"). Deep learning solves this problem by learning a sequence of non-linear transformations that result in feature representations that are better suited to classification. Such learned features have been shown to drastically outperform hand-engineered features in computer vision and audio analysis domains. However, applying the deep learning approach to the task of MRI classification is extremely challenging, because it requires a very large amount of data which is currently not available. We propose to instead use a three dimensional scattering transform, which resembles a deep convolutional neural network but has no learnable parameters. Furthermore, the scattering transform linearizes diffeomorphisms (due to e.g. residual anatomical variability in MRI scans), making the different disease states more easily separable using a linear classifier. In experiments on brain morphometry in Alzheimer's disease, and on white matter microstructural damage in HIV, scattering representations are shown to be highly effective for the task of disease classification. For instance, in semi-supervised learning of progressive versus stable MCI, we reach an accuracy of 82.7%. We also present a visualization method to highlight areas that provide evidence for or against a certain class, both on an individual and group level.

1. Introduction

[1] Over the last two decades, Magnetic Resonance Imaging (MRI) has been widely adopted for studying the human brain and diseases affecting brain tissue. The neurodegenerative processes behind these diseases are still poorly understood. Previous studies show complex and multivariate patterns of tissue damage, such as in Alzheimer's disease (Jack et al., 2004; Dyrba et al., 2015; Zhang et al., 2011; Cuingnet et al., 2011; Young et al., 2013; Moradi et al., 2015; Arbabshirani et al., 2016) and brain damage induced by HIV-infection (Su et al., 2016), compared to healthy controls.

Machine learning techniques have been proven powerful in identifying such multivariate patterns in an approach to classify patients and controls. Deep learning is a machine learning methodology currently transforming fields such as speech recognition (Hinton et al., 2012; Chorowski et al., 2015; Bengio and Heigold, 2014), image analysis (Xu et al., 2014), natural language processing (Pennington et al., 2014), high energy physics (Baldi et al., 2014), among others. Data in these domains usually look like low dimensional measurements where neighboring elements are highly correlated. Convolutional neural networks exploit this correlation and scale to massive datasets.

The situation in classifying MRI data is starkly different however. Here we collect data with millions of features, but the number of patients often does not exceed a few hundred or a few thousand. Applying high capacity deep learning methods in healthcare applications is therefore highly challenging and prone to severe overfitting.

In this work, we will show that deep models can be applied successfully to the medical imaging domain by applying a fixed

* Corresponding author.
E-mail address: tameem.hesham@gmail.com (T. Adel).

[1] Most of the changes performed in this resubmission are colored blue. However, in cases, like figures and equations, where we thought that having two different colors might cause confusion, color remains black.

http://dx.doi.org/10.1016/j.nicl.2017.02.004

(i.e. not learned) feature transformation to the input data. In detail, we apply a 3-dimension (3D) translation invariant transformation referred to as a 3D scattering transform, which is also stable to actions of small diffeomorphisms, such as deformations, to MRI data. In order to compute the scattering coefficients, a convolution network is established by cascading wavelet transforms and modulus operators (Bruna and Mallat, 2011; Mallat, 2012). The resulting 3D scattering representation of each data instance (MRI scan in our experiments), i.e. a vector containing multi-scale and multi-direction co-occurrence information, is subsequently used for classification. The performed experiments aim to show that the 3D scattering representation has more discriminative power than the original data features and than features derived from independent component analysis (ICA), and that it performs well in the low data regime. Effectively, by avoiding learning these transformations altogether, we also avoid overfitting.

Our main three contributions are as follows: First, we provide the first implementation of the feature representation referred to here as the 3D scattering transform, which is inspired by the 2D scattering transform proposed by Bruna and Mallat (2011), Mallat (2012). Second, we provide state-of-the-art performance on the classification in three datasets. These include the comorbidity and aging with HIV (AGEhIV) study looking into long-term effects of Human Immunodeficiency Virus (HIV)-infection on the brain, and two studies into (progression to) Alzheimer's disease (AD), namely the Open Access Series of Imaging Studies (OASIS) study (Marcus et al., 2007) and Alzheimer's Disease for Neuroimaging (ADNI) study (http://adni.loni.usc.edu). In the AD-studies, we will study gray matter volume, segmented in T1-weighted scans, while in the HIV-study we will assess white matter microstructure, measured with diffusion weighted MRI. Within the ADNI study, we will in more detail classify sub-classes with baseline data on Mild Cognitive Impairment (MCI) that will either progress into AD or remain stable. Third, we present a visualization method to highlight regions in the original MRI input that provide input for or against a certain class. This visualization method facilitates the understanding of how a particular classification decision is taken through the complex non-linear function resulting from the scattering transformation. Our method can visualize evidence on both the group and the individual level, and will be demonstrated on the ADNI-dataset.

The rest of the paper is organized as follows. Section 3 describes the 3D scattering transform, the relevant theoretical properties of this transform and the proposed visualization method. Empirical results on three MRI datasets on brain damage due to HIV (AGEhIV) and Alzheimer's disease (OASIS and ADNI) are presented in Section 4, followed by a conclusion.

2. The Scattering representation

The introduced paradigm consists of two principal steps. The first is to transform all data instances from their MRI representation into the scattering representation by the 3D scattering transform. The second is to perform supervised (respectively semi-supervised in the last experiment) learning. A scheme of the phases of the paradigm is displayed in Fig. 1. In the beginning, an initial procedure is performed to regress out the effect of age on the MRI scans. After estimating the age effect in a regression on the healthy controls, it is then regressed out of the data belonging to all classes. The 3D scattering transform is subsequently applied to all the age regressed MRI data, resulting in the scattering representation of each MRI dataset. A detailed description of the 3D scattering transform is provided throughout the rest of this section. Afterwards, learning is applied to the data in the scattering representation. Cross-validation is performed in all the experiments where a portion of 60% of the data is reserved for training, 20% for validation and 20% for testing. Supervised learning is performed using a linear support vector machines (SVM) classifier. The learning algorithm used in the last experiment, which is a semi-supervised learning experiment performed on the ADNI dataset, is a Laplacian SVM. Finally, we will present our method of visualization of evidence for or against a class, on both the group and individual level.

2.1. The 3D wavelet transform

We will first review the wavelet transform, before describing in detail the structure of the scattering transform. The 3D wavelet transform (Mallat, 1999) expands a three-dimensional signal on a basis of rotated and dilated wavelets. Wavelets are constructed from a Gabor mother wavelet ψ,

$$\psi(x) = \exp\left(-\frac{x^T \Lambda x}{2\sigma^2} + i\xi x\right), \tag{1}$$

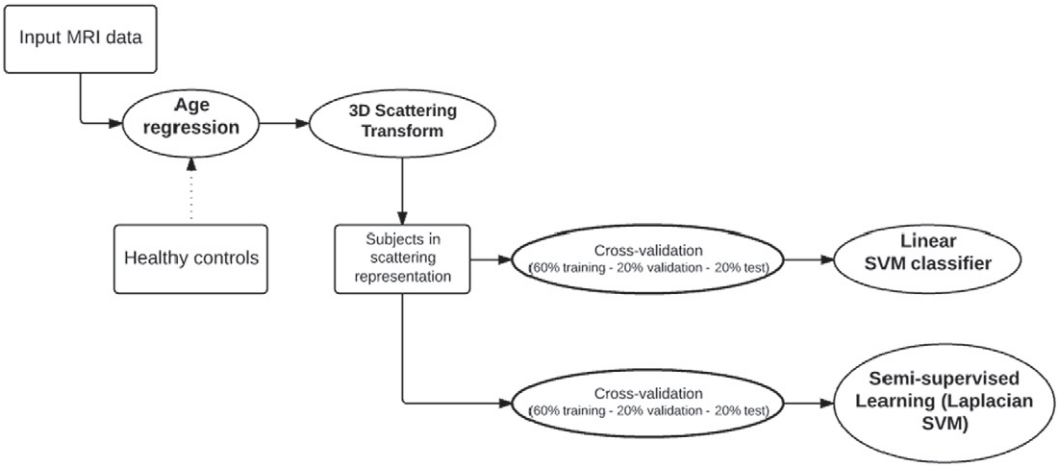

Fig. 1. A scheme of the main steps of the 3D scattering paradigm. A procedure to regress out the effect of age on the MRI scans, is performed in the beginning. The 3D scattering transform is subsequently applied to each MRI subject resulting in the scattering representation of the MRI data. Cross-validation is performed in all the experiments where data are split as follows: 60% training + 20% validation + 20% test. Supervised learning is performed using a linear support vector machines (SVM) classifier. The learning algorithm used in the last experiment, which is a semi-supervised learning experiment performed on the ADNI dataset, is a Laplacian SVM. Not shown is the introduced method of visualization of evidence for or against a class, on both the group and individual level, see Section 2.5.

where Λ is an optional diagonal metric matrix that controls the aspect-ratio of the filter and $x \in R^3$. The Gaussian window of the complex Gabor wavelets is modulated by a frequency, $\xi \in \mathbb{R}^3$, which controls the frequency of the filter relative to the width σ of the Gaussian window. We use $\xi = \left(\frac{3\pi}{4}, 0, 0\right)$.

A wavelet $\psi_{j,\theta,\beta,\gamma}$ at scale j and angles θ, β and γ is constructed from the mother wavelet by rotation and dilation:

$$\psi_{j,\theta,\beta,\gamma}(x) = \frac{1}{a^j\sigma}\psi\left(\frac{R_{\theta,\beta,\gamma}x}{a^j}\right) \tag{2}$$

where $R_{\theta,\beta,\gamma}$ is a rotation matrix and a is a scale parameter. $R_{\theta,\beta,\gamma}$ is a matrix used to perform rotation in a 3D space with angles, θ, β and γ in directions x, y and z, respectively, assuming x-y-z 3D coordinates (Arfken, 1985).

For notational convenience we define $\lambda = (j, \theta, \beta, \gamma)$ and write $\psi_\lambda = \psi_{j,\theta,\beta,\gamma}$.

The wavelet transform of a signal $f : \mathbb{R}^3 \to \mathbb{R}$ (e.g. an MRI scan) is then computed as

$$[Wf](\lambda, x) = f * \psi_\lambda(x). \tag{3}$$

where $*$ denotes convolution.

2.2. The scattering transform

The 3D scattering transform is computed as a sequence of multi-directional and multi-scale wavelet transforms, interleaved with modulus nonlinearities. The wavelet we use is a Gabor wavelet. The resulting computation is similar to a convolutional neural network, but does not have any learnable parameters, making it suitable for use in the low-data regime. At the first layer (layer 0), a Gaussian filter $\phi_J(x)$ at scale J is applied leading to a blurred version of x (see Fig. 2). The averages of these subsampled coefficients are then stored as part of the scattering representation. Then, a wavelet transform is applied to the MRI scan f: the convolutions $f * \psi_{j_1,\theta_1,\beta_1,\gamma_1}$ are computed for all j_1, θ_1, β_1 and γ_1 in a fixed grid (see Fig. 2) before storing the averages of these coefficients as another part of the scattering representation. After computing the modulus, we again blur and subsample (see Fig. 2), this time at scale j_q (j_q is the scale of last wavelet applied, as we will indicate later, this is equivalent to j_2 in our case and in the classification problems in general), and then store averages of the resulting coefficients as part of the scattering representation. A final wavelet transform is applied, yielding $|f * \psi_{j_1,\theta_1,\beta_1,\gamma_1}| * \psi_{j_2,\theta_2,\beta_2,\gamma_2}$ for all combinations of $j_1, j_2, (\theta_1, \beta_1, \gamma_1), (\theta_2, \beta_2, \gamma_2)$ where $j_2 < j_1$. After a final blurring and subsampling, we obtain the coefficients for the last layer.

The scattering transform of f consists of the computed coefficients $\mathcal{S}f$ at each layer:

$$\mathcal{S}f = \{f * \phi_J, \{|f * \psi_{\lambda_1}| * \phi_J\}_{\lambda_1}, \{||f * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi_J\}_{\lambda_1,\lambda_2}\} \tag{4}$$

Scattering coefficients can distinguish complex image structures, e.g. corners and junctions, from edges as they yield co-occurrence information for any pair of scales j_1 and j_2, and directions $(\theta_1, \beta_1, \gamma_1)$ and $(\theta_2, \beta_2, \gamma_2)$, where the subscript number indicates the order. Scattering coefficients are calculated only for scales $j_2 < j_1$, because, as proven in Mallat (2012), the energy in $|f * \psi_{j_1,\theta_1,\beta_1,\gamma_1}| * \psi_{j_2,\theta_2,\beta_2,\gamma_2}$ is negligible at scales $j_2 \geq j_1$.

2.3. The number of coefficients in a 3D scattering transform

Dependent on the number of scales and angular resolution, the scattering transform can greatly increase the dimensionality of the data. Hence, care must be taken in choosing the resolution settings for the scattering transform, as well as the regularization parameters. In this section we derive how the number of scattering coefficients depends on the settings, and describe the settings used in our experiments.

After averaging scattering coefficients, there is one scattering coefficient at order 0, which is the average of the original instance representation. The number of averaged scattering coefficients at orders 1 and 2 (most of the energy needed for classification lies in coefficients of orders 0, 1 and 2) depend on the number of wavelet scales J, and the resolution along each rotation axis θ, β, γ that we denote by L. At layer 1, there are $d_1 = JL^3$ wavelets and hence equally many feature maps and output coefficients. For each of the JL^3 feature maps, the second layer wavelet transform is computed, but only for smaller scales. Hence, the number of coefficients at layer 2 is

$$d_2 = L^3 \sum_{j=1}^{J} (J - j)L^3 = \frac{L^6(J-1)J}{2} \tag{5}$$

For example, for $L = 2, J = 4$, number of wavelet coefficients at orders 1 and 2 can be calculated as follows:

$$\begin{aligned} d_1 &= 4 \times 2^3 = 32 \\ d_2 &= 2^6 \times 3 \times 4/2 = 384. \end{aligned} \tag{6}$$

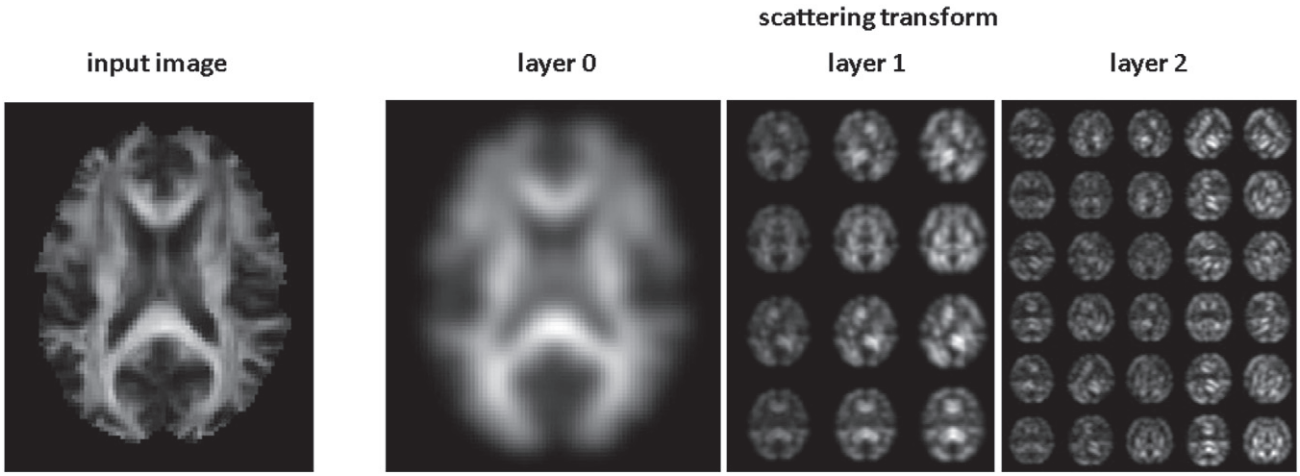

Fig. 2. A 2D slice of an MRI scan, and its representations at the first, second and third scattering layers (layers 0, 1 and 2).

Thus, the scattering transform representation of the original 3D MRI volume has dimensionality 417.

Similarly for $L = 4, J = 4$, the number of scattering coefficients at orders 1 and 2 can be calculated as follows:

$$d_1 = 4 \times 4^3 = 256$$
$$d_2 = 4^6 \times 3 \times 4/2 = 24576. \quad (7)$$

Thus, input is transformed to approximately $2.5 \cdot 10^4$ dimensions in this case. Assuming a resolution of 2 mm, the input dimensionality of an MRI-scan is approximately 10^6. In this situation, the dimensionality would therefore be reduced by roughly two orders of magnitude.

2.4. Linearization of diffeomorphisms

In brain morphometry studies, the differences between healthy and diseased states, and between diseased sub-states, are assumed to be described by non-linear deformations of the 3D volumetric data, which are mapped to modulated tissue probability maps that are to be compared.

An important step in the preprocessing is to remove any unwanted intensity variation, for instance due to anatomical misalignment between subjects by non-rigid registration. Such preprocessing will however not be able to remove all anatomical variations, such that residual misalignments will still be present in the data. Hence, it is important that the representation is stable to the actions of small diffeomorphisms (which explain most of the intra-class variability), while making the classes linearly separable.

As shown in Bruna and Mallat (2013), Mallat (2012), the scattering transform is Lipschitz continuous to deformations. This means that for a signal f, a diffeomorphism τ acting on the signal as $f_\tau(x) = f(x - \tau(x))$, there is a constant C such that:

$$\| Sf_\tau - Sf \| \leq C \| f \| \sup_x |\nabla \tau(x)|. \quad (8)$$

Here S denotes the scattering transform and $\sup x|\nabla\tau(x)|$ measures the magnitude of the deformation via the norm of the deformation gradient, $\nabla\tau(x)$ (Bruna and Mallat, 2013). This result shows that the difference between the scattering transform of f and its transformed version f_τ is bounded by a constant C times the norm of f times the magnitude of the deformation.

It follows from the Lipschitz continuity that the effect of a small diffeomorphism on the scattering representation is well-approximated by a linear map. As we verify in our experiments, this ensures that linear classifiers are able to handle variability due to deformations well.

The idea of the Lipschitz continuity in the context of the scattering transform was first noted and illustrated in pages 2, 3 and 4 -starting from Eq.(2)- in Mallat (2012). The idea of the Lipschitz continuity in general can as well be found in Chapter 9 in O'Searcoid (2006) .

2.5. Visualization of the 3D scattering transform

We propose a visualization method via which regions in the MRI scans providing evidence for or against a certain class can be highlighted. As such, the proposed visualization method provides insight into the decision making process of a linear classifier acting on the scattering representations of input MRIs. The visualization method is based on the gradient of the learning function with respect to the input features of an input MRI scan. We thus aim to quantify the contribution of voxel values per individual subject to the classification boundary. These are furthermore averaged per class (patients or controls), resulting in distinct class mappings. Importantly, this approach differs from the commonly visualized classifier coefficients. To reduce noise effects in computing the gradients, we compute gradients by perturbing principal components (PCs) of the input data.

More formally, for an input data sample, x, and a discrete classification output (class), y, let the the learning function be $G(x) = P(y|x)$, h_k be the PCs of the input matrix (of the original voxels), k be indices over the selected PCs, and ϵ_k be small values used to perturb h_k. Using Taylor's expansion, we can write:

$$G(x + \epsilon_k h_k) = G(x) + \epsilon_k \, (\nabla G(x) \cdot h_k) + o(\epsilon^2) \quad (9)$$

$$\frac{G(x + \epsilon_k h_k) - G(x)}{\epsilon_k} \approx \nabla G(x) \cdot h_k, \quad (10)$$

since we can safely neglect terms of $o(\epsilon^2)$. Also note that h_k forms an orthonormal basis, for which the following holds:

$$\nabla G(x) = \sum_k (\nabla G(x) \cdot h_k) h_k \quad (11)$$

Based on Eqs. (10) and (11), we get:

$$\nabla G(x) \approx \sum_{k=1 \cdots K} \frac{G(x + \epsilon_k h_k) - G(x)}{\epsilon_k} h_k \quad (12)$$

And this can be the case with small values of ϵ. Eq. (12) can be approximated by using our PCs h_k. Recall that these are computed in the original voxel space, and not over the scattering operators. Thus, by perturbing h_k we approximately compute the gradient of $G(x)$ by Eq. (12) and obtain the directions of maximum change in $G(x)$. The directions of maximum change for each input image are the weights of the respective principal directions. The number of PCs can be selected as a small fraction of the total number of included subjects in the dataset. Using fewer PCs leads to more blurring.

The learning model, which in our case is an SVM classifier on top of the scattering representation, is expressed in this visualization methodology by $G(x)$. Using probabilities is one common way of expressing $G(x)$, but is not necessary. We use a function of the distance from the corresponding instance, x, to the boundary of the SVM, divided by the sum of the distances of all training instances belonging to the same class, to the boundary, to compute $G(x) = P(y|x)$. This is valid since the farther an instance is from the boundary, the more certain it is that this is the right class for such instance. An alternative is to use Platt scaling, which is calibrated but it requires post-processing and a separate validation set.

3. Experiments

In our experiments, we will compare three different feature representations:

1. The original voxel data, reshaped as 1D vectors.
2. An independent component analysis (ICA) based representation, as implemented in the Group ICA for fMRI toolbox (GIFT) (Calhoun et al., 2009) after an initial Principal Component Analysis (PCA) mapping retaining 90% of the variance in the data.
3. The scattering representation.

These will be computed on three brain MRI datasets, that will be explained in more detail in the following sections:

1. The AGEhIV (comorbidities and aging in HIV) dataset, including scans of HIV-infected individuals and healthy controls. In this study, data have been collected on two MRI scanners.

Before classifying disease type, we will perform a sub-experiment to identify the source scanner used to record the MRI scans.

2. The OASIS (Open Access Series of Imaging Studies) dataset including patients with Alzheimer's disease (AD) and healthy controls.
3. The ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset including patients with Alzheimer's disease, (different states of) Mild Cognitive Impairment (MCI) and healthy controls. This dataset is considerably larger than the OASIS dataset. This allows us to study in more detail the transitional stage between age-related cognitive decline and AD (Moradi et al., 2015) referred to as mild cognitive impairment (MCI). One experiment depicts the supervised learning (classification) task of discriminating between two different MCI states, referred to as progressive and stable MCI (pMCI and sMCI). The last ADNI experiment represents a semi-supervised learning task where MRI scans that are known to belong to the MCI label, but not known to which substate of MCI they belong, are used as the unlabeled sample along with samples belonging to pMCI and sMCI. Adding these data is thought to improve the classification (Moradi et al., 2015).

The classifier in use is an SVM with a soft margin. SVM is one of the most commonly used classification algorithms and it was demonstrated in previous works that it is more accurate than alternatives, e.g. Othman et al. (2011), Yang et al. (2011). The dual form of soft margin SVMs (13), which is formulated into a quadratic programming problem, is solved using sequential minimal optimization (SMO) (Platt, 1999),

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) \\ &\text{w.r.t.} \ \ \alpha_i, \quad \text{subject} \quad \text{to}: \ 0 \le \alpha_i \le C, \ \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{13}$$

where α_i for $i = 1\ldots,n$ denote Lagrange multipliers, and $\alpha_i = 0$ for all training instances except the support vectors. The symbol K denotes the SVM kernel.

Linear SVM (SVM with a linear kernel) is used in all the supervised learning experiments. Cross-validation is also applied to all the performed experiments. Training is performed on only 60% of the available data, whereas 20% of the data is reserved for validation of the scattering parameters, namely number of orientations, L, and number of scales, J (respectively validation of number of independent components in case of the ICA representation), among other parameters. The test set consists of the remaining 20% of the data.

3.1. AGEhIV dataset

The AGEhIV Cohort Study is an ongoing study on prevalence, incidence and risk factors of ageing-associated comorbidities and organ dysfunction among HIV patients and highly comparable HIV-uninfected controls of at least 45 years of age (i.e. same geographic region with similar socio-demographic and behavioral(risk) factors) (Schouten et al., 2014).

The AGEhIV dataset consists of MRI data of 100 HIV-infected patients and 60 healthy controls, included in the Academic Medical Center (AMC) in Amsterdam, The Netherlands. Of these subjects, diffusion weighted MRI data were acquired at 2 mm resolution. Preprocessing of the data was performed with software developed in-house (Matlab, MathWorks, Natick, MA), using the HPCN-UvA Neuroscience Gateway and using resources of the Dutch e-Science Grid (Shahand et al., 2014). As a result, Fractional Anisotropy (FA) maps were computed. FA is sensitive to microstructural damage and therefore expected to be, on average, decreased in patients. Subjects were scanned on two 3.0 Tesla scanner systems, 121 subjects on a Philips Intera system and 39 on a Philips Ingenia system. Patients and controls were evenly distributed over both scanners. More details on the study, data and preprocessing can be found elsewhere (Su et al., 2016).

We perform two experiments on the AGEhIV dataset. In the main experiment, Experiment A, the class labels refer to the disease type, i.e. being diagnosed with or without HIV. The used MRI scanner is added as an additional feature to the voxel data. In Experiment B, MRI scanner is used as the class label, and HIV-status is added as a feature to the data. In these experiments we chose not to linearly regress out age, but rather add age as an additional feature to the voxel data, similar to the above-mentioned covariates.

The number of ICA components was chosen by optimizing performance on the validation set and resulted in 9 components. For the scattering representation, parameters leading to the results displayed in Table 1 have also been determined on the validation set. The angular resolution is $L = 4$, while the number of wavelet scales used is $J = 4$, which leads to a total of 24,833 features, i.e. scattering coefficients.

Table 1 shows the SVM classification results of both experiments A and B. In experiment A, SVM applied to the scattering representation of the MRI scans achieves better accuracy than SVM over the raw voxel form, and than SVM over the ICA-based representation. Adding up age and scanner information does not improve the overall classification accuracy. An equivalent improvement for the scattering representation compared to the other two representations is obtained in experiment B.

As mentioned in Section 2.3, most of the energy required to capture discriminative information for classification lies in scattering coefficients of orders 0, 1 and 2. Therefore, all results obtained here are based on two scattering layers. To validate this choice, we performed an additional experiment using one, two and three layers of

Table 1
Accuracy obtained by applying an SVM classifier to the MRI original features (voxels), ICA-based representation and scattering representation of the HIV dataset.

Features	Accuracy	Standard dev.	Standard error
Experiment A: classifying HIV-status			
Original (91 × 109 × 91)	67.3%	2.4	0.6
ICA (9 components)	68.2%	2.8	0.7
Scattering (24,833)	75.1%	2.4	0.6
Scattering (24,833) + age + scanner	75%	2.2	0.6
Experiment B: classifying scanner			
Original (91 × 109 × 91) + age + HIV-status	79.6%	2.8	0.7
ICA (9 components) + age + HIV-status	76.1%	3	0.8
Scattering (24,833) + age + HIV-status	84.8%	2.6	0.7

coefficients. The classification accuracies obtained were 58.0%, 75.1% and 75.0% respectively. Thus, involving scattering coefficients of order 3 leads to a massive increase in the number of used scattering coefficients without any performance gain. Both of these conclusions have as well been reached by Bruna and Mallat (2011), Bruna (2013).

Fig. 3 (a) displays a histogram showing the difference in mean error, i.e. error of SVM applied to the original features minus the error of SVM applied to the scattering features, for different data partitions. A positive difference thus refers to a larger error of SVM applied to the original features.

Receiver operating characteristic (ROC) curves for the SVM classifier applied to the three representations of the HIV dataset are shown in Fig. 4 (a). For the classification results arising from the ICA components vs. the scattering representation, the p-value for a paired Student's t-test is equal to $p = 0.045$, which means that the null hypothesis – being that the pairwise difference in the results between the two representations is not significant – is rejected at a standard significance level of $p = 0.05$. The p-value of the raw form (original voxels) vs. the scattering representation is equal to $p = 4.48 \times 10^{-4}$.

Out of the 24,833 scattering features, 12,483 features have values smaller than 10^{-14}, across all instances (MRI scans). By discarding such features and selecting the rest to be the input to SVM, identical accuracy results are obtained with a sped up run-time. We used a linear kernel for the results reported in Table 1. The training set consists of 60 HIV and 36 healthy scans. The validation set consists of 20 HIV as well as 12 healthy scans, and the same goes for the test set.

3.2. OASIS Alzheimer dataset

The Open Access Series of Imaging Studies (OASIS) dataset consists of high-resolution T1-weighted MRI scans of 182 MRI scans, including 60 Alzheimer's patients and 122 healthy controls. These data were preprocessed using the Statistical Parametric Mapping software (SPM v12). This involved tissue segmentation into gray matter and white matter probability maps, followed by spatial normalization to a pre-defined standard template, in MNI152 space. The registration parameters for the normalization were calculated using the DARTEL non-linear approach in SPM. The non-linear warping was then applied to the gray matter and white matter segmentations separately to align them to the template, during which images were resampled to 1.5 mm isotropic voxels, modulated by the extent of warping necessary in order to retain volumetric information and smoothed using a 4 mm FWHM Gaussian kernel. The resulting volumes contain ($120 \times 144 \times 120$) voxels.

Scattering representations of the 3D scans are used as the input to an SVM classifier. MRI scans are then classified into one of two classes: Alzheimer or healthy.

As mentioned above, the raw form representation of each OASIS MRI scan contains $120 \times 144 \times 120$ features. 13 ICA components are used in the ICA-based representation experiment, which was determined on a validation set. Again, an angular resolution, $L = 4$, and number of wavelet scales, $J = 4$, were used to extract the scattering coefficients.

Experimental results on the OASIS data are displayed in Table 2. Out of the three representations, SVM applied to the scattering representation of the MRI scans achieves the highest accuracy.

Fig. 3 (b) displays a histogram showing the difference in mean error between SVM applied to the original features and SVM applied to the scattering features, i.e. error of SVM applied to the original features - error of SVM applied to the scattering features. Again, a positive difference refers to a larger error of SVM applied to the original features.

Receiver operating characteristic (ROC) curves for the SVM classifier applied to the three representations of the dataset are shown in Fig. 4 (b). For the classification results of the ICA-based representation vs. the scattering representation, the p-value of a paired T-test is equal to 0.06. The p-value for the classification accuracy of SVM applied to the raw MRI form vs. SVM applied to the scattering representation is 0.05.

Similar to Section 3.1, a simple feature selection mechanism is applied to the scattering coefficients where coefficients with value $< 10^{-14}$ are removed for the sake of optimizing computational runtime. The training set consists of 36 Alzheimer and 74 healthy scans. The validation set consists of 12 Alzheimer as well as 24 healthy scans, same size as the test set.

3.3. ADNI database

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The

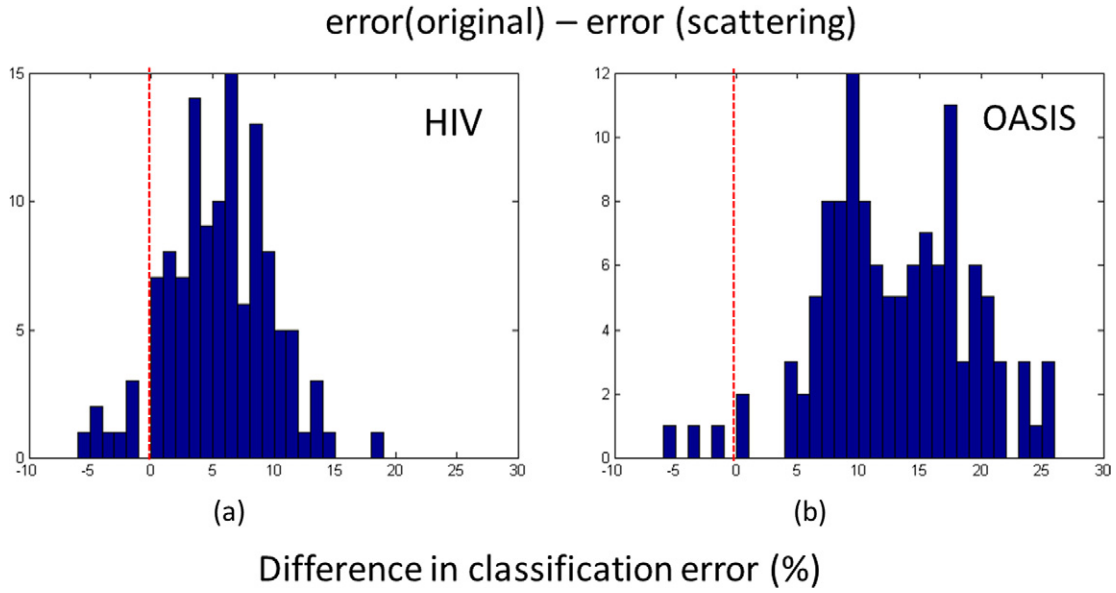

Fig. 3. Histograms of the error of SVM applied to the original features minus the error of SVM applied to the scattering features. A positive difference refers to a larger error of SVM applied to the original features, (a) for the HIV dataset, (b): for the OASIS dataset.

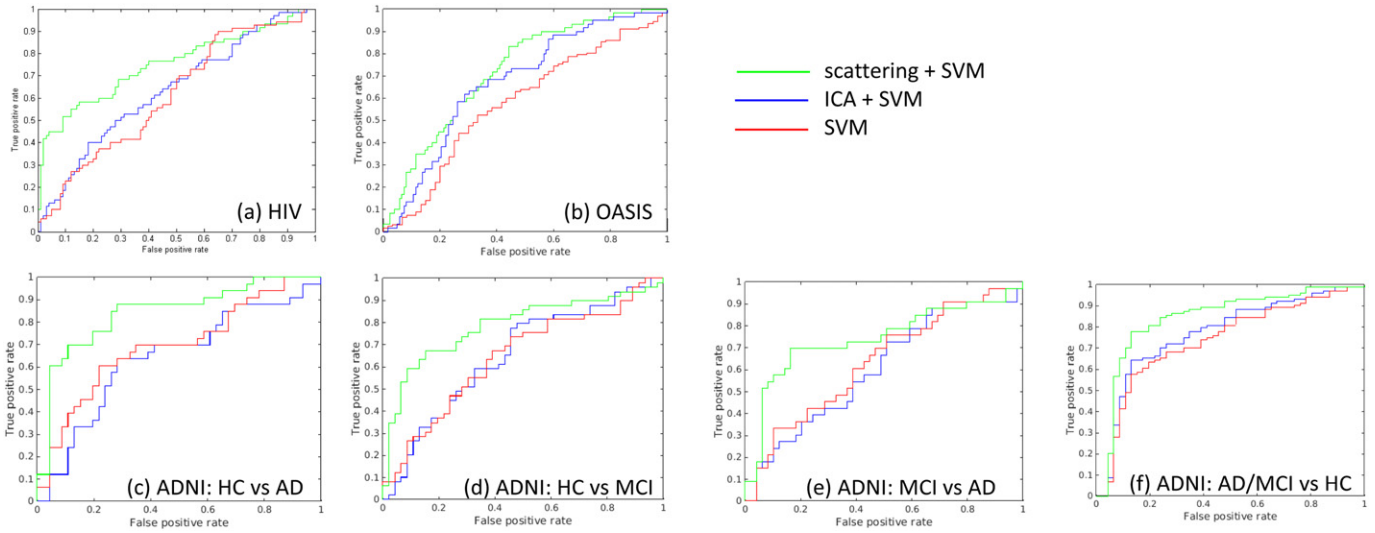

Fig. 4. ROC curves of the SVM classifier applied to the raw form, ICA and scattering representation of the: (a): HIV dataset. (b): OASIS dataset. (c–f): ADNI subset of 150 subjects.

primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), 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). For up-to-date information, see www.adni-info.org. We conduct experiments on two subsets of the ADNI dataset, one, which stands in line with the main genesis of the work, consists of a relatively small sample of each class. On the other hand, the other sample is a bigger sample by which we assess supervised and semi-supervised learning performances based on the scattering feature representation in cases where the training data are not too small.

3.3.1. Experiments on a subset of data

Focussing on the main theme of this work, which is to assess the scattering transform in small datasets, we initially work with a subset of 150 MRI scans consisting of 50 AD scans, 50 MCI scans and 50 healthy scans. The subjects were randomly chosen from each class, and they cover different sex and acquisition differences (recall that age has already been regressed out). Similar to previous works on the same dataset, we perform binary classification experiments consisting of all possible two-class combinations. We also perform an experiment combining AD and MCI scans (we refer to this composite class as "disordered") against healthy scans.

Each ADNI MRI scan used in the experiments is a normalized $160 \times 192 \times 160$ voxel volume. In the ICA-based representation experiment, 18 ICA components are used, as determined on a validation set. Again, an angular resolution, $L = 4$, and number of wavelet scales, $J = 4$, are used to extract the scattering coefficients. The training set consists of 30 AD, 30 MCI and 30 healthy scans. The validation set consists of 10 AD, 10 MCI and 10 healthy scans, and the same goes for the test set.

Table 2
Accuracy obtained by applying an SVM classifier to the MRI original features (voxels), ICA-based representation and scattering representation of the OASIS dataset.

Features	Accuracy	Standard dev.	Standard error
OASIS: classifying Alzheimer's diseases and healthy controls			
Original ($120 \times 144 \times 120$)	62.2%	2.8	0.7
ICA (13 components)	66.4%	2.4	0.6
Scattering (24,833)	73%	2.6	0.7

Experimental results on the ADNI dataset are displayed in Table 3. Out of the three representations, SVM applied to the MRI scattering representation achieves significantly better performance than the raw form and the ICA-based representations.

The four ROC-curves correspondent to the four experiments in the upper part of Table 3 are shown in Fig. 4 (c–f). Again, all experiments are SVM classification experiments applied to the three representations of the ADNI dataset. As can be seen in the figures, classification results of SVM over the scattering representation are more accurate than the raw form and than the ICA-based representation.

The p-values of a paired T-test for the ADNI experiments are displayed in Table 4.

3.3.2. Experiments on a bigger sample size

The larger sample of ADNI used in this experiment consists of 835 MRI scans. One of the advantages of using this ADNI sample in particular is that - contrary to the majority of experiments performed on the ADNI dataset in the literature - it allows us to compare on common ground with a previous work on the data, since the same sample was used in the experiments performed in Moradi et al. (2015). This sample consists of 200 Alzheimer's disease (AD) MRI scans, 231 healthy MRI scans and 404 MCI MRI scans. We perform two classification tasks and a semi-supervised learning task on this ADNI sample. The first is a binary classification task of the AD vs. healthy MRI scans. Two MCI-based tasks are performed on the MCI subset of this ADNI sample, one classification and one semi-supervised learning task. The goal of the MCI classification task is to discriminate between progressive MCI (pMCI) scans and stable MCI (sMCI) scans. In this task, the goal is to predict whether an MCI patient will convert to an AD over a 3-year period (referred to as pMCI) or not (referred to as sMCI) (Moradi et al., 2015). The MCI MRI scans we have in this sample include 164 pMCI and 100 sMCI scans besides other MCI scans not known as to which MCI state they belong (unlabeled MCI scans).

Regarding the classification of the AD and healthy MRI scans, the subsample used for this task is composed of 200 AD and 231 healthy MRI scans. The number of ICA components in use in the ICA-based representation experiment is 17, as determined on a validation set. As in the corresponding AGEhIV, OASIS and small ADNI sample experiments, an angular resolution, $L = 4$, and number of wavelet scales, $J = 4$, are used to extract the scattering coefficients. A

Table 3
Accuracy obtained by applying an SVM classifier to the MRI original features (voxels), ICA-based representation and scattering representation of the ADNI dataset.

Experiment	Feature rep.	Accuracy	Standard dev.	Standard error
ADNI: experiments on a subset of 150 subjects				
AD vs. healthy	Original (160 × 192 × 160)	63.3%	3.2	0.8
AD vs. healthy	ICA	67%	2.8	0.7
AD vs. healthy	Scattering (24,833)	78.5%	3.2	0.8
MCI vs. healthy	Original (160 × 192 × 160)	66%	3	0.8
MCI vs. healthy	ICA	66.6%	3.6	0.9
MCI vs. healthy	Scattering (24,833)	79.4%	3	0.8
AD vs. MCI	Original (160 × 192 × 160)	63.9%	3.6	0.9
AD vs. MCI	ICA	67.3%	2.8	0.7
AD vs. MCI	Scattering (24,833)	72.2%	3	0.8
(AD + MCI) vs. healthy	Original (160 × 192 × 160)	66.8%	2.8	0.7
(AD + MCI) vs. healthy	ICA	69%	2.8	0.7
(AD + MCI) vs. healthy	Scattering (24,833)	83.8%	3	0.8
ADNI: experiments on a larger set of 835 subjects (Moradi et al., 2015)				
AD vs. healthy	Original	65.8%	4	1.0
AD vs. healthy	ICA	70.9%	3.6	0.9
AD vs. healthy	Scattering (24,833 coefficients)	84.9%	3.7	0.9
AD vs. healthy	Scattering (20 PCs)	86.4%	3.4	0.9
AD vs. healthy	Moradi et al. (2015)	82%	–	–

proportion of 60% of the data is used for training, and 20% of the data is used for validation and for testing. Results of the AD vs. healthy classification experiment are shown in the lower part of Table 3. The classification accuracy achieved by the proposed scattering-based classifier is 84.9%, compared to state-of-the-art leading to corresponding accuracy of 82% by Moradi et al. (2015). The scattering-based classifier leads to state-of-the-art results on this sample of ADNI, better than the other two representations, and better than the result reported in Moradi et al. (2015). We perform a single additional experiment to reduce the number of scattering coefficients by PCA. The first 20 principal components (PCs) were selected and given as an input to the SVM classifier, instead of the scattering features. This results not only in speeding up the classifier, but also in an improvement in the classification accuracy, as can be seen in Table 3.

Results of the ADNI MCI experiments are displayed in Table 5. Data for the classification experiment consist of 164 pMCI and 100 sMCI MRI scans. By applying a linear SVM classifier to the scattering coefficients of these scans, a classification accuracy of 73.5% is achieved, based on cross-validation (upper half of Table 5), compared to state-of-the-art accuracy on the same data ranging from 66.0% to 69.2%, as reported in Table 3 in Moradi et al. (2015).

Data for the semi-supervised learning (SSL) experiment consisted of 164 pMCI, 100 sMCI and 140 unlabeled MCI scans. The SSL algorithm we use on top of the scattering representation is a variation of semi-supervised SVM (Melacci and Belkin, 2011) where Laplacian SVM is efficiently trained in the primal with preconditioned conjugate gradient based on the L_2 hinge loss[2] . The learning accuracy resulting from using Laplacian SVM on the scattering coefficient representation is superior to the best SSL classification accuracy reported in Table 3 in Moradi et al. (2015). It can also be noticed that adding unlabeled MRI data to data of both algorithms improves the learning accuracy over their fully supervised learning counterparts.

3.4. Scattering visualization

We present a visualization of the classification of scattering coefficients on the supervised learning experiment of pMCI vs. sMCI only. Based on Eq. (12), gradients of all MRI scans belonging to both the pMCI and sMCI classes are computed.

[2] Code for Laplacian SVM used in our experiments is available at http://sourceforge.net/projects/lapsvmp/

Fig. 5 shows an example on two single pMCI and two sMCI brains, with minimal and maximal distance to the classification boundary. Sub-cortical regions receive higher weight when classifying pMCI, while cortical gray matter provides more evidence for classifying sMCI. The cingulate cortex provides evidence for both classes.

Additionally, averaging those maps over subjects per class provides more generic information. Of interest, in the stable MCI group, larger cortical gray matter areas receive a high weight. In the progressive MCI group, we see effects in the hippocampus, amygdala, entorhinal cortex, precuneus and cingulate cortex.

4. Discussion and conclusion

In this paper, we propose a 3D scattering transform as a fixed representation with no learnable parameters, that can replace deep learning methods in paradigms where we do not have at disposal massive amounts of data points available for use. We present a visualization method of evidence for belonging to a certain class, both for individual cases and class averages. In our experiments we have extensively validated our method on different datasets, showing improved performance over currently established SVMs, also with ICA added as a dimensionality reduction step.

The scattering transform and deep learning are directly related (see Bruna and Mallat (2011, 2013), Mallat (2012) for an in depth explanation). For instance, there is a pooling operation in scattering networks implemented as a subsampling step and there is also a nonlinearity implemented as the absolute value of the activities (instead of a rectified linear unit (ReLU)). The fact that features are extracted from every layer also has an equivalent in Convolutional

Table 4
P-values of a paired T-test for the four ADNI experiments on the scattering representation vs. original and ICA-based representations.

ADNI: comparing performance on different feature representations		
Experiment	Feature representation	*P*-value
AD vs. healthy	Scattering vs. Original	0.0012
AD vs. healthy	Scattering vs. ICA	0.002
MCI vs. healthy	Scattering vs. Original	0.0025
MCI vs. healthy	Scattering vs. ICA	0.02
AD vs. MCI	Scattering vs. Original	0.04
AD vs. MCI	Scattering vs. ICA	0.048
(AD + MCI) vs. healthy	Scattering vs. Original	0.0007
(AD + MCI) vs. healthy	Scattering vs. ICA	0.001

Table 5
Results of the MCI experiments. The first experiment is a supervised learning experiment with two labels, pMCI and sMCI. Using a linear SVM on top of the scattering representation leads to a superior classification accuracy compared to state-of-the-art by Moradi et al. (2015). The second experiment is an SSL experiment where the training data consist of unlabeled MCI MRI data in addition to data belonging to the pMCI and sMCI labels. Using a Laplacian SVM trained in the primal on top of the scattering representation leads to a higher accuracy than state-of-the-art by Moradi et al. (2015).

Experiment	Algorithm	Accuracy	Standard dev.	Standard error
ADNI MCI: supervised and semi-supervised learning.				
Supervised	Scattering + SVM	73.5%	3.7	0.9
	Moradi	69.2%	–	–
Semi-supervised	Scattering + Laplacian SVM	82.7%	3.2	0.8
	Moradi SSL	74.7%	–	–

Neural Networks (CNNs) called "skip-connections" (Graves, 2013; He et al., 2016; Huang et al., 2016). Finally, the SVM is not all that different from the final fully connected logistic regression layers in CNNs, since both are linear classifiers. Since in our case where we do not train the parameters of the CNN (it is given by a fixed scattering transform) the SVM is more convenient because software packages exist that very quickly and reliably optimize the SVM's parameters. Also, very good software packages for semi-supervised learning exist for the SVM. Next we describe CNNs before moving to the principal conclusions of this work.

Convolutional neural networks (CNNs) are a type of feed-forward artificial neural networks. CNNs convolve the input image (or voxel grid) by a set of learned filters, resulting in a so-called *feature map* for each filter. Each response value (neuron activation) in a feature map is the result of a small filter operating on a small region of the input referred to as the receptive field of the neuron. CNNs achieve insensitivity to small translations and deformations by pooling activations in small regions of the feature maps, for example by computing the maximum over a 2×2 region (max-pooling) or average (mean-pooling). Deep networks can be constructed by stacking convolutions, pooling operators and point-wise nonlinearities such as rectified linear activation functions. CNNs have been intensively applied to vision and image recognition problems (Farabet et al., 2010; Matusugu et al., 2013; Ciresan et al., 2013; Behnke, 2003; Yaniv et al., 2015; Masci et al., 2013), document recognition (LeCun et al., 1998) and medical signal processing (Graupe et al., 1988, 1989), among numerous other applications. However, the learned CNN transformations have mostly led to very good classification results in cases where a lot of data is available. Since learning the accompanying parameters necessitates the availability of significant amounts of data, a high risk of overfitting occurs when using deep learning techniques in low-data regimes.

The proposed 3D scattering representation is a fixed representation with no learnable parameters, which is one of the reasons why it is much less prone to severe overfitting than CNNs in low-data regimes. The scattering transform is translation invariant and

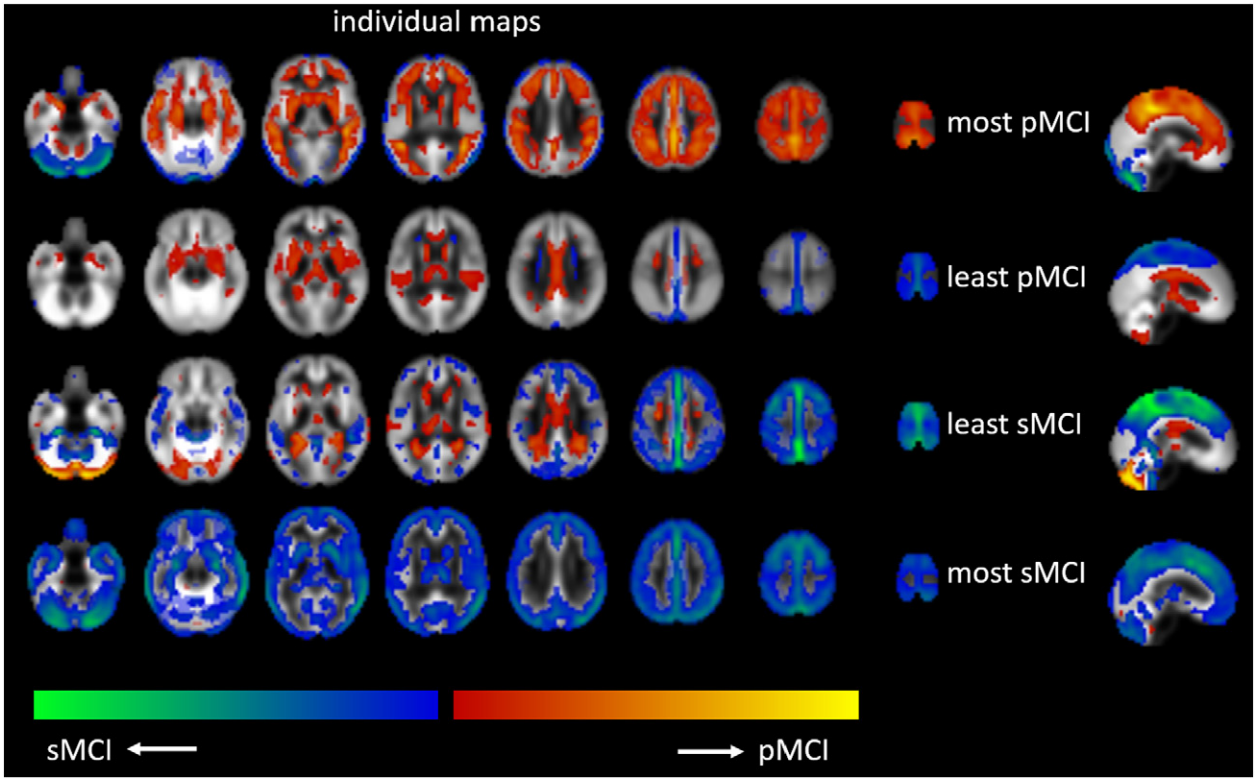

Fig. 5. Visualization of evidence for four individual participants of the ADNI study to be assigned to progressive and stable MCI (pMCI/sMCI) classes. Participants that were classified as most or least probable pMCI/sMCI were selected. Red/yellow voxels show evidence for the pMCI class, i.e. increases in the values of the red/yellow voxels lead to the scan being more likely to be a pMCI scan. Vice versa, blue/green voxels show evidence for the sMCI class. Maps are thresholded on 20% of the maximum value.

stable to small deformations. The scattering transform involves convolutions followed by an averaging step (can as well be seen as a mean-pooling step), and a supervised learner (classifier) is applied on top of it so that labels can be learned. As demonstrated by the experiments, the premise is that the scattering transform can be a better representation for the data in terms of explaining the important variations and discarding the rather unimportant variations, and this ultimately leads to better classification accuracy.

We have developed and implemented a feature extraction method based on three dimensional scattering transformations and tested it for its ability to discriminate HIV and Alzheimer's disease from healthy subjects and subjects with mild cognitive impairment. We have clearly shown that our proposed methodology achieves higher accuracy than the best competing methods, in particular an SVM applied to features extracted using ICA. We believe that one of the main reasons that scattering is successful in the neuroimaging domain is its stability against small deformations of the input image. Also its ability to detect differences in "tissue textures" may be important. We believe scattering representations are particularly useful in the high-dimensional but small-number-of-patients regimes that are typical in medical imaging.

In our experiments, we have used two open source databases. In Yang et al. (2011), a fast ICA representation was implemented and used for the classification of OASIS MRI signals. It achieved an average accuracy of 70.7%, which is outperformed by the 73% classification performance of our scattering representation. Regarding the ADNI-study, the average MCI vs. Healthy classification accuracy over the fast ICA representation implemented in Yang et al. (2011) is 72%, compared to a classification accuracy of 79.4% achieved by the scattering representation. For the AD vs. Healthy experiment, the classification accuracy reported in Yang et al. (2011) is 76.9%, whereas the corresponding scattering classification accuracy is 78.5%. The scattering representation outperforms Fast ICA in spite of the fact that we use only 50 MRI scans per class, in contrast with 202 AD, 410 MCI and 236 healthy MRI scans used in the experiments conducted in Yang et al. (2011). The work in Gupta et al. (2013), which is based on a high-data regime where a sparse auto-encoder was used to learn a set of bases from natural images, uses 755 ADNI MRI scans per class, compared to our 50 scans per class in our experiments, and they consequently achieve higher classification results. In the context of classifying progressive MCI, Moradi et al. (2015) showed that they classified beyond the state of the art. Here we show that the proposed scattering transform allows to further improve upon the classification. Still, our classification performance is not an outlier compared to earlier studies (Arbabshirani et al., 2016), reducing the likelihood of possible overtraining. In the ADNI-dataset age effects were regressed out on voxel-level in a univariate way, whereas for the AGEhIV dataset we followed an alternative approach by adding nuisance variables as independent features. Thus, we have carried out different but equally valid approaches to handle this issue throughout the different performed experiments.

Visualizing the classification result is of high clinical importance. In the current practice, a classification experiment is interpreted by studying the classifiers' weight vector. In this respect, previous work studied the interpretation of weight vectors in backward models (Haufe et al., 2013), and provided an analytical approximation to permutation testing for running computationally efficient experiments (Gaonkar and Davatzikos, 2013).

We propose a method that allows to study the probability of a single subjects' brain to be assigned to a specific class. When averaging over the progressive MCI group, as can be seen in Fig. 6, we observed that the hippocampus, amygdala and entorhinal cortex were involved. These regions have been reported earlier in the context of predicting progression in MCI (Ye et al., 2012)

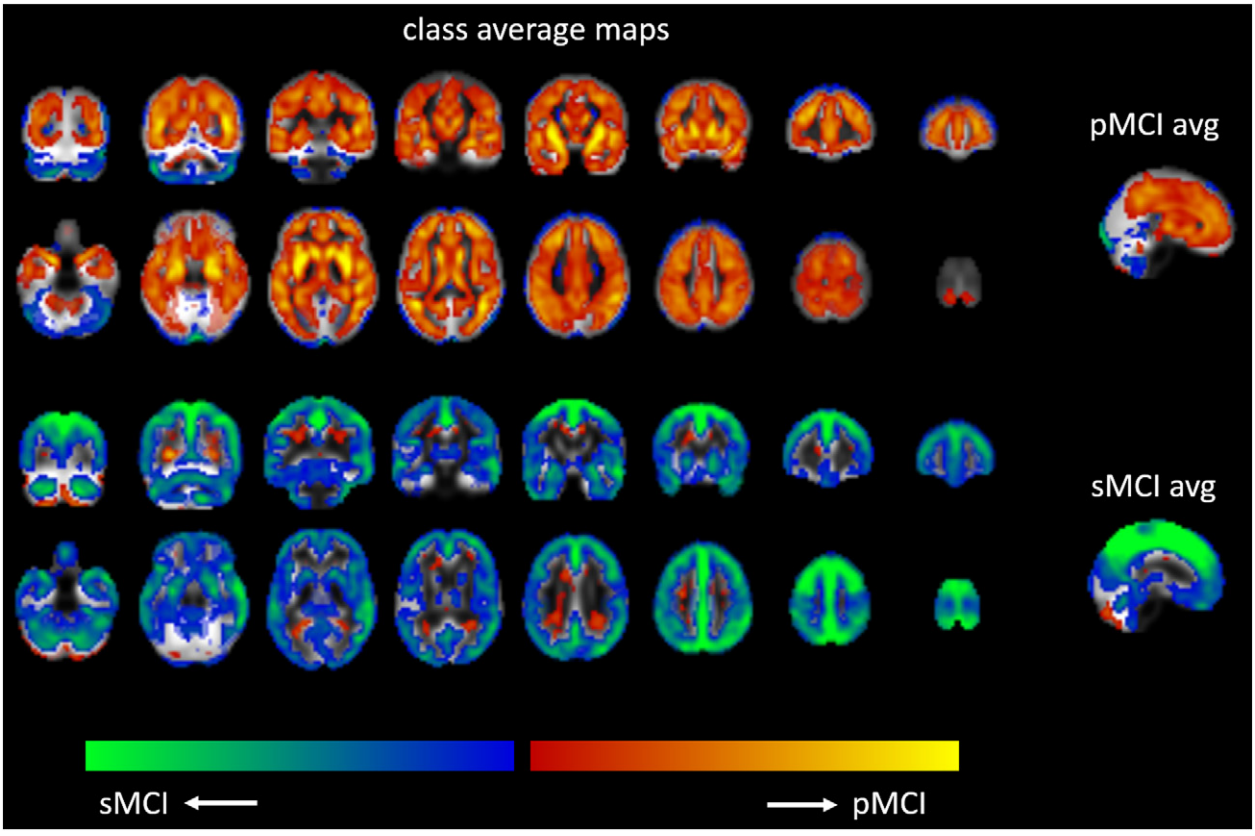

Fig. 6. Visualization of evidence for pMCI and sMCI classes, averaged over all participants per class.

and discriminating MCI from healthy controls (Desikan et al., 2009). Atrophy in the enthorinal cortex has been shown to predict cognitive decline in Alzheimer's disease (Velayudhan et al., 2013). Hippocampal atrophy has been observed in earlier stages of Alzheimer research (Henneman et al., 2009). An additional strong effect that we observe in the precuneus and cingulate cortex has been shown to be associated to hypometabolism in MCI (Bailly et al., 2015).

Although we have not done so, we believe that regression problems could also benefit from applying the scattering transform to the data. This would provide a more natural embedding of regressing out covariates such as age and scanner. We have done this in a separate step and a univariate way, which is a limitation of our experiments.

In conclusion, we propose a scattering transform that proved to be highly effective in small datasets, under different experimental conditions and for multiple disease types. The classification can be visualized on both the individual and group level.

Acknowledgment

We gratefully acknowledge support from the NWO IPPSI-KIEM program no.: 628.005.012.

We warmly thank Dr. James Cole for providing us with preprocessed OASIS data, and Dr. Jussi Tohka for providing us with the preprocessed ADNI data of their paper Moradi et al. (2015).

The work described in this paper is partially performed using the AMC Neuroscience Gateway, using resources of the Dutch e-Science Grid with the support of SURF Foundation. The acquisition of the AGEhIV HIV-data was supported by the Nuts-OHRA Foundation (grant no. 1003-026), Amsterdam, The Netherlands, as well as by The Netherlands Organisation for Health Research and Development (ZonMW) together with AIDS Fonds (grant nos. 300020007 and 2009063, respectively). Additional unrestricted scientific grants were received from Gilead Sciences, ViiV Healthcare, Janssen Pharmaceutica N.V., Bristol-Myers Squibb, Boehringer Ingelheim, and Merck&Co. None of these funding bodies had a role in the design or conduct of the study, the analysis and interpretation of the results, or the decision to publish.

The OASIS dataset is supported by grant numbers P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

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.; 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Arbabshirani, M., Plis, S., Sui, J., Calhoun, V., 2016. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage.
Arfken, G., 1985. Mathematical Methods for Physicists. 3rd ed., Academic Press.
Bailly, M., Destrieux, C., Hommet, C., Mondon, K., Cottier, J., Beaufils, E., Vierron, E., Vercouillie, J., Ibazizene, M., Voisin, T., Payoux, P., 2015. Precuneus and cingulate cortex atrophy and hypometabolism in patients with Alzheimer's disease and mild cognitive impairment: MRI and 18 f-FDG PET quantitative analysis using freesurfer. Biomed. Res. Int. 1–8.
Baldi, P., Sadowski, P., Whiteson, D., 2014. Searching for exotic particles in high-energy physics with deep learning. Nat. Commun. 5,
Behnke, S., 2003. Hierarchical Neural Networks for Image Interpretation. Springer Science & Business Media.
Bengio, S., Heigold, G., 2014. Word embeddings for speech recognition. Int. Speech Commun. Assoc. 15,
Bruna, J., 2013. Scattering Representations for Recognition. Dissertation Ecole Polytechnique X.
Bruna, J., Mallat, S., 2011. Classification with scattering operators. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR).
Bruna, J., Mallat, S., 2013. Invariant scattering convolution networks. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 35 (8), 1872–1886.
Calhoun, V., Liu, J., Adali, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. Neuroimage 45, 163–172.
Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. Adv. Neural Inf. Proces. Syst. (NIPS).
Ciresan, D., Meier, U., Masci, J., Gambardella, L., Schmidhuber, J., 2013. Flexible, high performance convolutional neural networks for image classification. Proc. Twenty-Second Int. Joint Conf. Artif. Intell. (AISTATS) 22, 1237–1242.
Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 62, 591–600.
Desikan, R., Cabral, H., Hess, C., Dillon, W., Glastonbury, C., Weiner, M., Schmansky, N., Greve, D., Salat, D., Buckner, R., Fischl, B., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimers disease. Brain 132,
Dyrba, M., Grothe, M., Kriste, T., Teipel, S., 2015. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. Hum. Brain Mapp. 36, 2118–2131.
Farabet, C., Kavukcuoglu, K., LeCun, Y., 2010. Convolutional networks and applications in vision. Proc. Int. Conf. Circuits Sys. (ISCAS).
Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage 78, 270–283.
Graupe, D., Liu, R., Moschytz, G., 1988. Applications of neural networks to medical signal processing. IEEE Conf. Decis. Control 27, 343–347.
Graupe, D., Vern, B., Gruener, G., Field, A., Huang, Q., 1989. Decomposition of surface EMG signals into single fiber action potentials by means of neural network. Proc. Int. Conf. Circuits Sys. (ISCAS) 1008–1011.
Graves, A., 2013. Generating Sequences with Recurrent Neural Networks. arXiv:1308.0850
Gupta, A., Ayhan, M., Maida, A., 2013. Natural image bases to represent neuroimaging data. Proc. Int. Conf. Mach. Learn. (ICML) 30, 987–994.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J., Blankertz, B., Bießmann, F., 2013. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87C, 96–110.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. Eur. Conf. Comput. Vis. (ECCV).
Henneman, W., Sluimer, J., Barnes, J., Flier, W.V.D., Sluimer, I., Fox, N., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. Neurology 72 (11), 999–1007.
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Vanhoucke, A.S.V., Nguyen, P., Sainath, T., Kingsbur, B., 2012. Deep neural networks for acoustic modeling in speech recognition. Signal Process. Mag. 6, 82–97.
Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K., 2016. Deep networks with stochastic depth. arXiv:1603.09382.
Jack, C., Shiung, M., Gunter, J., O'brien, P., Weigand, S., Knopman, D., Boeve, B., Ivnik, R., Smith, G., Cha, R., Tangalos, E., 2004. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 62, 591–600.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.
Mallat, S., 1999. A Wavelet Tour of Signal Processing. Academic Press.
Mallat, S., 2012. Group invariant scattering. Commun. Pure Appl. Math. 65 (10), 1331–1398.
Masci, J., Giusti, A., Ciresan, D., Fricout, G., Schmidhuber, J., 2013. A fast learning algorithm for image segmentation with max-pooling convolutional networks. IEEE Int. Conf. Image Process. (ICIP) 20, 2713–2717.
Matusugu, M., Katsuhiko, M., Yusuke, M., Kaneda, Y., 2013. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw. 16, 555–559.

Melacci, S., Belkin, M., 2011. Laplacian support vector machines trained in the primal. J. Mach. Learn. Res. (JMLR) 12, 1149–1184.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.
Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 1498–1507.
O'Searcoid, M., 2006. Metric Spaces. Springer Science & Business Media.
Othman, M., Abdullah, N., Kamal, N., 2011. MRI brain classification using support vector machine. Model. Simul. Appl. Optimization (ICMSAO) 4,
Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. Empir. Methods Nat. Lang. Process. (EMNLP) 12, 1532–1543.
Platt, J., 1999. Fast training of support vector machines using sequential minimal optimization. Adv. kernel Methods 185–208.
Schouten, J., Wit, F., Stolte, I., Kootstra, N., Van Der Valk, M., Geerlings, S., Prins, M., Reiss, P., 2014. Cross-sectional comparison of the prevalence of age-associated comorbidities and their risk factors between hiv-infected and uninfected individuals: the age HIV cohort study. Clin. Infect. Dis. 59, 1787–1797.
Shahand, S., Benabdelkader, A., Jaghoori, M., AlMourabit, M., Huguet, J., Caan, M., van Kampen, A., Olabarriaga, S., 2014. A Data-centric Neuroscience Gateway: Design, Implementation, and Experiences. Concurrency and Computation: Practice and Experience.
Su, T., Caan, M., Wit, F., Schouten, J., Geurtsen, J., Cole, J., Sharp, J., Vos, F., Prins, M., Portegies, P., Reiss, P., Majoie, C., Charles, B., 2016. White matter structure alterations in HIV-1-infected men with sustained suppression of viraemia on treatment. AIDS 30, 311–322.
Velayudhan, L., Proitsi, P., Westman, E., Muehlboeck, J., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Spenger, C., Hodges, A., 2013. Entorhinal cortex thickness predicts cognitive decline in Alzheimer's disease. J. Alzheimers Dis. 33, 755–766.
Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Chang, I., 2014. Deep learning of feature representation with multiple instance learning for medical image analysis. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) 1626–1630.
Yang, W., Lui, R., Gao, J., Chan, T., Yau, S., Sperling, R., Huang, X., 2011. Independent component analysis-based classification of Alzheimer's disease MRI data. J. Alzheimers Dis. 24, 775–783.
Yaniv, B., Diamant, I., Greenspan, H., 2015. Deep learning with non-medical training used for chest pathology identification. SPIE Med. Imaging.
Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Victor, N., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V., 2012. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol. 12 (1).
Young, J., Modat, M., Cardoso, M., Mendelson, A., Cash, D., Ourselin, S., 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage Clin. 2, 735–745.
Zhang, D., Wang, Y., Yuan, L.Z.H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867.