Medical Image Analysis 44 (2018) 228–244

Contents lists available at ScienceDirect

Medical Image Analysis

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

Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation

Sérgio Pereira[a,b,1,*], Raphael Meier[c,1], Richard McKinley[d], Roland Wiest[d], Victor Alves[b], Carlos A. Silva[a], Mauricio Reyes[c]

[a] CMEMS-UMinho Research Unit, University of Minho, Guimarães, Portugal
[b] Centro Algoritmi, University of Minho, Braga, Portugal
[c] Institute for Surgical Technology and Biomechanics, University of Bern, Switzerland
[d] Support Center for Advanced Neuroimaging – Institute for Diagnostic and Interventional Neuroradiology, University Hospital and University of Bern, Switzerland

ARTICLE INFO

Article history:
Received 28 March 2017
Revised 15 October 2017
Accepted 12 December 2017
Available online 20 December 2017

Keywords:
Interpretability
Machine learning
Representation learning

ABSTRACT

Machine learning systems are achieving better performances at the cost of becoming increasingly complex. However, because of that, they become less interpretable, which may cause some distrust by the end-user of the system. This is especially important as these systems are pervasively being introduced to critical domains, such as the medical field. Representation Learning techniques are general methods for automatic feature computation. Nevertheless, these techniques are regarded as uninterpretable "black boxes". In this paper, we propose a methodology to enhance the interpretability of automatically extracted machine learning features. The proposed system is composed of a Restricted Boltzmann Machine for unsupervised feature learning, and a Random Forest classifier, which are combined to jointly consider existing correlations between imaging data, features, and target variables. We define two levels of interpretation: global and local. The former is devoted to understanding if the system learned the relevant relations in the data correctly, while the later is focused on predictions performed on a voxel- and patient-level. In addition, we propose a novel feature importance strategy that considers both imaging data and target variables, and we demonstrate the ability of the approach to leverage the interpretability of the obtained representation for the task at hand. We evaluated the proposed methodology in brain tumor segmentation and penumbra estimation in ischemic stroke lesions. We show the ability of the proposed methodology to unveil information regarding relationships between imaging modalities and extracted features and their usefulness for the task at hand. In both clinical scenarios, we demonstrate that the proposed methodology enhances the interpretability of automatically learned features, highlighting specific learning patterns that resemble how an expert extracts relevant data from medical images.

1. Introduction

Machine learning approaches can be broadly divided into two categories: those using hand-crafted features, and those relying on Representation Learning techniques. Representation Learning refers to a set of general machine learning methods for automatic learning and extraction of features directly from data. By contrast, hand-crafted features require expert knowledge on the problem, hence making them more problem-dependent (LeCun et al., 2015). Notwithstanding, there is usually a data representation mapping stage that takes the input data and transforms it into a more discriminative representation.

Despite the success of Representation Learning-based methods (Salakhutdinov et al., 2007; Krizhevsky et al., 2012; Pereira et al., 2016a; Kamnitsas et al., 2016), they are often regarded as uninterpretable "black boxes". This is due to the large number of layers, or nodes, which makes it difficult to unveil the relations between inputs and outputs. In fact, this undesirable characteristic is shared with other machine learning models, such as Random Forests (RFs) with many trees, or linear models with thousands of features

* Corresponding author at: Department of Electronics Campus Azurém, Guimarães, Portugal.
E-mail addresses: id5692@alunos.uminho.pt (S. Pereira), raphael.meier@istb.unibe.ch (R. Meier), RichardIain.McKinley@insel.ch (R. McKinley), roland.wiest@insel.ch (R. Wiest), valves@di.uminho.pt (V. Alves), csilva@dei.uminho.pt (C.A. Silva), mauricio.reyes@istb.unibe.ch (M. Reyes).
[1] Shared first author.

https://doi.org/10.1016/j.media.2017.12.009
1361-8415/

(Ribeiro et al., 2016a; 2016b; Lipton, 2016). Thus, in general, there is a trade-off between the capacity/complexity of the model and its interpretability. Nevertheless, while much of the focus on machine learning has been dedicated to solve complex problems with high performance, the possibility of interpreting a decision of a model is still a very desirable property of a predictive system, but has not received much attention so far. This is especially important with the pervasive adoption of machine learning-based models in critical areas such as in radiology (Wang and Summers, 2012), where a prediction should not be blindly followed. “Black box” models may look untrustworthy in the sense that a decision cannot be explained, especially in case of failure.

However, as stated by Lipton (2016), trust and interpretability may be ill-defined. We can understand trust as the confidence in the model itself, or its prediction (Ribeiro et al., 2016b; Freitas, 2014). In the former, we can trust a model if we have confidence that it will behave as expected after deployment. We can base this trust on the measured performance. Yet, adversarial examples (Szegedy et al., 2014; Nguyen et al., 2015) show us that unpredictable, or bizarre, behaviors may arise, even in highly accurate systems. Trusting a prediction means that we have enough confidence that a given prediction is correct (Ribeiro et al., 2016a; 2016b; Lipton, 2016). Interpretability is a way to enhance trust in a system. Understanding the predictions and how information is encoded in a model can help us to comprehend as to why it fails, and avoid the undesirable trial and error development procedure (Zeiler and Fergus, 2014). Interpretability can appear as an interpretation of the model itself, or as a post-hoc interpretation of the model through its predictions. The former is hindered by the complexity of the model, which is usually proportional to the difficulty of the task and model performance. On the other hand, post-hoc interpretability is based on a qualitative explanation of an already trained system, by means of visualization, or study of examples. In this way, there is less need to sacrifice model’s performance/complexity for the sake of interpretability (Ribeiro et al., 2016a; 2016b; Lipton, 2016). To further contextualize these ideas, the state of the art in model interpretation is presented in the following section.

1.1. Previous work

We can broadly differentiate between two approaches for model interpretation: *global* and *local* interpretation. The global interpretation of a machine learning model is aimed at understanding *how* information extracted from the input data is used by the model to perform predictions. Local interpretation is aimed at understanding *why* a certain decision was made by the model at hand.

In order to enable global interpretability, some authors proposed to simplify, or transform the models (Tibshirani, 1996; Olden and Jackson, 2002; Craven and Shavlik, 1996; Hara and Hayashi, 2016). Tibshirani (1996) proposed Lasso to force some weights of the model to be exactly 0, which enhances its interpretability. Craven and Shavlik (1996) converted a previously trained neural network into a more interpretable decision tree. Olden and Jackson (2002) proposed a method to remove unimportant connections from a neural network. Hara and Hayashi (2016) interpreted tree ensembles by approximating a simpler model derived from the minimization of the KL-divergence between that simpler model and the more complex model. Gallego-Ortiz and Martel (2016) deduced rules from RFs and presented them for interpretation. These proposals have been defined for particular models and attempt to simplify models to make them globally interpretable. While these models provide information about how the model learned the training data, they can be less practical when there are high-dimensional feature vectors.

Another group of methods that are more model-agnostic treat the model as “black boxes”, and bring understanding about their decisions. This is accomplished either by perturbation of the features (Cortez and Embrechts, 2011; Krause et al., 2016; Ribeiro et al., 2016b), or by fitting a simpler model to the predictions of the more complex one (Baehrens et al., 2010; Ribeiro et al., 2016a; 2016b). These approaches are post-hoc in the sense that they do not explain the inner workings of the model itself, but its predictions. Particularly, perturbing features (Cortez and Embrechts, 2011; Krause et al., 2016; Ribeiro et al., 2016b) and observing its impact on the decision may provide an estimate of the feature importance, but it does not take into account correlations among features. Additionally, it may be impractical for high-dimensional feature vectors. Other approaches try to interpret the learning algorithm locally, i.e., its behavior in the vicinity of the test samples (Baehrens et al., 2010; Ribeiro et al., 2016a, 2016b). To this end, Baehrens et al. (2010) approximated the predictions of the model under analysis with another model. Then, an explanation vector was defined as the derivative of the probabilistic output in relation to the data point. Explanation vectors provide information about which features would affect more the prediction of that sample. Although the method by Baehrens et al. (2010) may provide some insight about the model, a human may not extract any interpretation from it, if the explanation vector is high-dimensional. Ribeiro et al. (2016b) also pursued local interpretability as a way to achieve model agnostic interpretation that could be applied even for very complex deep networks. For a given sample in the feature space of the model, a set of synthesized examples in the vicinity of the sample is created, whose prediction is obtained from the model under analysis. Then, a simpler model is fitted to these samples and interpreted. The simpler model does not represent the original complex model globally, but it is an approximation of its behavior in the vicinity of the given sample. Nevertheless, this approach is agnostic to the model being interpreted and robust to its complexity. Most of the previous proposals relied on some sort of visualization to present the data for human interpretation, focusing on approximations of the model under analysis. Other approaches purely relied on visualizing the topology of neural networks (Hinton et al., 1986; Wejchert and Tesauro, 1989; Tzeng and Ma, 2005). Zrihem et al. (2016) used t-SNE (Maaten and Hinton, 2008) in the context of deep reinforcement learning to reduce the dimensionality of neural activations of a Deep Q Network, in order to study the policies of the agent at hand. Visualization in the context of Convolutional Neural Networks (CNN) comes in the form of saliency maps that inform which region of the image was important for a given class (Simonyan et al., 2013), or deconvolving an activation and projecting it in the image space (Zeiler and Fergus, 2014). The later involves coupling a deconvolutional neural network to the CNN under analysis.

High-dimensional feature vectors increase the computational load and complexity of a machine learning model, as well as the risk of overfitting due to irrelevant features having spurious correlations with the target variable. Furthermore, it may render the interpretability of a model more difficult (Tibshirani, 1996). Hence, feature selection may be seen as a prerequisite for enabling model interpretation. Univariate feature selection methods evaluate the relationship of each feature with some condition of interest, but cannot detect interactions among features, which is the advantage of multivariate methods. Some of the latter approaches are wrappers around a learner that iteratively evaluate subsets of features in relation to their predictive power (Ganz et al., 2015). However, these recursive feature elimination methods may be unpractical for very large datasets. Random Forests also stand as a multivariate approach for feature selection (Konukoglu and Ganz, 2014), due to their capability to measure feature importance through the mean decrease impurity (MDI), allowing us to rank features. The draw-

back is that one still needs to choose a user-defined threshold on the ranking, or, as proposed by Konukoglu and Ganz (2014), on an upper bound on false positive rates in selecting unimportant features as relevant ones.

Taking into account the aforementioned interpretability-related studies, one can draw some conclusions. 1) Post-hoc approaches provide tools to potentially interpret more complex models. This copes with modern trends favoring powerful, yet complex, approaches, such as methods based on deep learning. 2) A model can be agnostically, but locally, interpreted, providing insights regarding each individual decision. This may allow us to infer about its coherence and to reason about mistakes. 3) On the other hand, approaches that focus on a global interpretation can provide clues about how the model learned to look at the data, however, they are model-specific and lack the local interpretability necessary to understand individual predictions. 4) High-dimensional feature representations may pose difficulties for interpretation. Thus, effective feature selection methods might be required. 5) Visualization tools are natural human understandable data exploration tools for enhancing interpretability. 6) Interpretation is ultimately performed by the human expert. For example, in radiology, clinical experts tend not to trust machine decisions as the interpretability and ultimately the trustworthiness of automatic algorithms tend to be low (Wang and Summers, 2012). If we want to increase trust, we should devise methodologies for interpretability that are simple and understandable by humans. Additionally, with interpretability methodologies, we expect to retrieve hints to answer the following questions: 1) How does a system use the input data to solve the task at hand? 2) When a system fails, why does it fail? 3) Is the system capturing the relevant relations in the data? For instance, Ribeiro et al. (2016b) found that a system that was accurately detecting Husky dogs in pictures was basing its prediction in the presence, or not, of snow and not in the dog itself. This is an example of a system that bases its predictions on a feature that does not coherently fit into the concept of the target variable (Husky dog).

1.2. Motivation and contributions

Motivated by the current trend for favoring complex models at the expense of interpretability, in this paper we propose to interpret a machine learning system both globally and locally. In contrast to previous studies, we hypothesize that both global and local interpretability provide complementary insights about the machine learning system at hand. The focus of this work is on the interpretation of automatically learned features for lesion segmentation in medical images. Particularly, we will study the interpretation of features stemming from Magnetic Resonance Imaging (MRI) sequences. We propose to drive feature selection for the task at hand by coupling a Restricted Boltzmann Machine-based data representation model with a RF classifier, to jointly consider existing correlations between imaging data, features, and target variables. The contributions in this work are the following: i) We explore a strategy for global interpretability by inferring which parts of the input data contributed the most for highly important features in different segmentation tasks, thus indirectly interpreting how a RBM encoded the data. ii) We interpret image segmentations locally through assessing the spatial relevance of the features distributed in the image space. Finally, iii) in order to leverage interpretability, we also evaluate a joint Mutual Information and RF feature importance strategy for automatically selecting important features. We evaluate the proposed approaches in brain tumor segmentation and penumbra estimation in ischemic stroke lesions. In an ischemic stroke, the closest tissues to the blocked blood vessels are at high risk of infarction. Those tissues around this core that suffer from the reduced blood supply, but are still salvageable, form the penumbra region. The databases used in this study for brain tumor segmentation and penumbra estimation are publicly available and being actively used in recent research (Menze et al., 2015; Maier et al., 2017); thus, enabling future comparison with our proposal.

The remainder of this paper is organized as follows. In Section 2, we introduce the two basic components used for our interpretable machine learning system. The proposed system, feature selection, and interpretability methodologies are presented in Section 3. In Section 4 we describe the databases and experimental setup. Results are presented in Section 5. Then, in Section 6, we discuss our results. Finally, we draw the main conclusions in Section 7.

2. Preliminaries

Our machine learning system is based on a Restricted Boltzmann Machine (RBM) (Smolensky, 1986) to learn features, and a RF classifier (Breiman, 2001) as discriminative model. RBMs are generative unsupervised Representation Learning techniques that learn the intrinsic representation of the data without information regarding any target variable (Hinton et al., 2006). Hence, RBMs can be trained on large unlabeled data sets, as typically found in the clinics. This represents an advantage over supervised learning approaches, as manual labeling of large data sets is expensive, time-consuming, and prone to intra- and inter-observer variability. Features extracted by these unsupervised models have proven to be useful for texture classification (van Tulder and de Bruijne, 2016) and volume estimation (Zhen et al., 2016). Since RBMs are an unsupervised method, after feature learning it is necessary to employ a task-related supervised learning algorithm (van Tulder and de Bruijne, 2016, 2015; Zhen et al., 2016), or supervised fine-tuning stage (Nair and Hinton, 2010) that learns how to map the learned features to the desired task. RFs are one of the possible supervised learning algorithms, which have shown advantages as a classifier, such as being robust to overfitting, and successfully dealing with high-dimensional feature vectors by identifying and ranking relevant features (Criminisi and Shotton, 2013). While this capability of RFs may be used to find which features better explain the target variables, or for feature selection (Konukoglu and Ganz, 2014), getting unbiased feature importance measures is not trivial (Louppe et al., 2013). Due to its advantages, RFs have been successfully used in medical image analysis, e.g. Criminisi and Shotton (2013); Meier et al. (2014a); Zhen et al. (2016); Menze et al. (2016); Pereira et al. (2016b); Maier et al. (2017); Meier et al. (2016); McKinley et al. (2016).

Since high-dimensional feature vectors may impair the interpretability of a model, feature selection is required (Tibshirani, 1996). Filter-based feature selection methods have the advantages of scaling well and being computationally efficient. Mutual Information (MI) is an information measure that can be used to assess feature relevance. Moreover, it has the advantage of measuring any kind of relation among variables, even non-linear ones (Bennasar et al., 2015; Vergara and Estévez, 2014). Methods based on this measure evaluate the MI between features and target variables (Peng et al., 2005; Battiti, 1994). However, there is no involvement of the learning algorithm that is supposed to employ the selected features in the actual selection procedure. In this paper, however, we use MI both between data and features, as well as between features and task-related classes through a RF.

3. Methods

In this work, we consider two main blocks: the machine learning system, and the interpretability system (Fig. 1). Taking into account the previous work (Section 1.1), feature selection is required

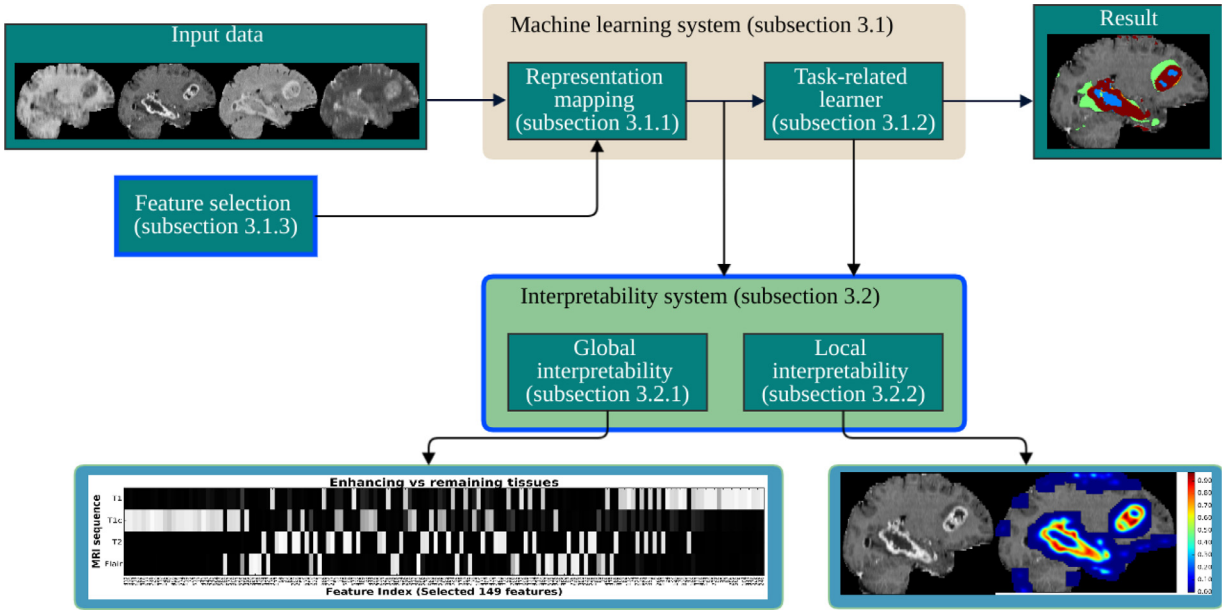


Fig. 1. Proposed system. The machine learning system is composed of a representation mapping stage that generates the input features for a task-related learner, which computes the prediction. Feature selection is performed to enable an effective interpretation of the machine learning system. In order to enhance model interpretability, the combined use of global and local interpretability is proposed. Blue colored frames mark the modules representative of the main contributions in this paper. The visualization of the training stage of the machine learning system and feature selection is omitted for simplicity. We show an example application in brain tumor segmentation.

to enable an effective model interpretation. Thus, we will first introduce our machine learning system and propose a methodology for feature selection. Subsequently, we will present methods for model interpretation, which exploit the feature selection.

3.1. Machine learning system

There are two main stages in a machine learning system: representation mapping and the task-related learner. The former corresponds to the feature computation stage, which can be performed by representation learning or feature engineering. The latter is the predictive model, which is task-dependent since it is a supervised learning algorithm.

3.1.1. Representation mapping

We use RBM (Smolensky, 1986) to realize the representation mapping stage of our machine learning system. This is an undirected graphical Representation Learning model. The nodes are organized into one visible and one hidden layer, whose states are represented by the vectors $\mathbf{v} = [v_i : i = 1, \ldots, m]$, and $\mathbf{h} = [h_j : j = 1, \ldots, n]$, respectively. All nodes in one layer are connected to all nodes in the other layer with weights represented by the matrix $\mathbf{W} = [w_{ij}]$. No intra-layer connections exist. In this work, the inputs to the visible layer are patches of shape $d \times d \times d$ extracted from the set of available MRI sequences $\mathcal{C}$. Then, the patches are represented as a 1D vector and fed into the visible layer; thus, $m = d \cdot d \cdot d \cdot |\mathcal{C}|$ (Fig. 2). Originally, RBMs were proposed to model binary data in both layers. However, image patches are represented by a continuous range of values. Thus, the visible units are defined as linear units with independent Gaussian noise, which allows us to model continuous-valued inputs (Hinton, 2012). Noisy Rectifier Linear Units (NReLU) are used to represent the hidden units, since they proved to be suitable for feature extraction (Nair and Hinton, 2010). Thus, after receiving the input in the visible layer, the RBM can compute the activations in the hidden layer, thus mapping the input into a feature vector. The joint configuration of the states of the visible and hidden units is represented by an energy function defined as

$$E(\mathbf{v}, \mathbf{h}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{1}$$

where a_i is the bias of the visible unit i, b_j is the bias of the hidden unit j, and σ_i represents the standard deviation of the Gaussian noise of v_i (Hinton, 2012; Nair and Hinton, 2010). Having the energy function, the joint probability distribution over $\mathbf{v}$ and $\mathbf{h}$ is

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})}, \tag{2}$$

where Z represents the partition function. Computing Z is impractical, but we can still sample in parallel the state of all the units in one layer conditioned on the other layer, given that there are no intra-layer connections. Thus, we sample the hidden and visible units, respectively, as (Nair and Hinton, 2010; van Tulder and de Bruijne, 2015).

$$P(h_j|\mathbf{v}) = \max\left(0, \sum_i w_{ij} v_i + b_j + \mathcal{N}\left(0, sigm\left(\sum_i w_{ij} v_i + b_j\right)\right)\right), \tag{3}$$

$$P(v_i|\mathbf{h}) = \mathcal{N}\left(\sum_j w_{ij} h_j + a_i, \sigma_i\right), \tag{4}$$

where $\mathcal{N}$ represents the Gaussian distribution and *sigm* the sigmoid function.

We use Contrastive Divergence (Hinton, 2002) with one step of alternating Gibbs sampling to train the model. Since learning σ is difficult, following Hinton (2012), we normalize each component of the data with zero mean and unit variance and consider $\sigma_i = 1$. We also employ momentum, and both L1 and L2 weight-decay. L1 enforces sparsity, which leverages interpretability (Hinton, 2012), similarly to Lasso (Tibshirani, 1996). After training, we compute features as noise-free activations of the NReLU units. These units exhibit intensity equivariance if they are noise free and have zero biases ($b_j = 0$) (Nair and Hinton, 2010).

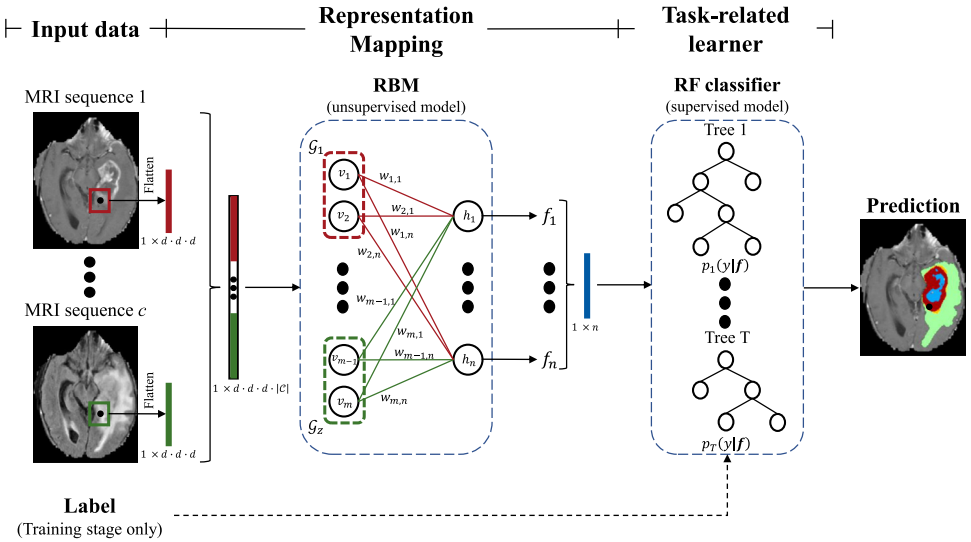


Fig. 2. Machine learning system. We use RBM as Representation Mapping and RF classifier as Task-related Learner. Patches are extracted from each MRI sequence, flattened, and concatenated into one single 1D vector. The RBM receives the imaging data in the visible layer, and maps it into a feature vector, as activations of the hidden units. $\mathcal{G}_y$ identify meaningful groups of visible units that receive data from a distinct MR sequence. The color in the connections identify weights that are linked to a given MRI sequence. The feature vector is fed into the RF classifier, which outputs a prediction for the central voxel of the patch (black dot).

3.1.2. Task-related learner

The RBM learns features from data in an unsupervised way, and without knowing the task for which the features will be used for. So, we need a supervised learning algorithm to learn how to make predictions out of those features. We use a RF classifier (Breiman, 2001) as task-related learner, by training it in a supervised way. This model is an ensemble of Decision Trees, each one trained on a randomly selected subset of the training set with replacement (bootstrap). Each training and testing sample is represented by the activations of the hidden layer of the RBM. So, each sample is represented by a n-dimensional feature vector and fed into the RF (Fig. 2). As the samples traverse the trees, a subset of features (randomly chosen during training) is evaluated in each node. This characteristic, together with the number of trees, allows the algorithm to deal with high-dimensional feature vectors. The randomness in the algorithm allows it to be robust to overfitting. At the same time, although RF can estimate the feature importance, getting unbiased measures is not trivial (Louppe et al., 2013; Criminisi and Shotton, 2013). For the reader interested in more details regarding the RF classifier, we refer to Breiman (2001) and Criminisi and Shotton (2013).

3.1.3. Joint RBM-RF mutual information approach for feature selection

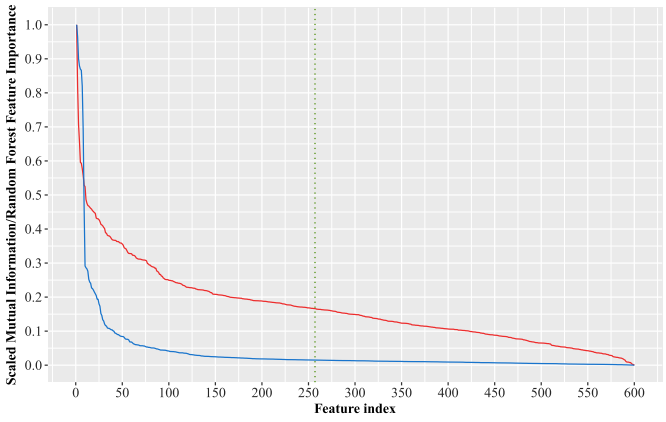
Usually, a feature represents the response of a feature detector applied over the data. Having noise is harmful since the learning algorithm may capture spurious correlations between the features and the labels. Additionally, feature noise, arising from detectors that enhance spurious variations in the data, may be adverse for the learning algorithm (Zhu and Wu, 2004). So, we hypothesize that a good feature should correlate with the class labels, and represent the data from which it was computed; hence, effectively connecting data with class labels. If it holds true, then interpreting a prediction may be more feasible in terms of which input caused it. Since MI is a measure of statistical dependence, we employ it to quantify the quality of the mapping between data and labels, through the features. Our procedure consists of the following steps: First, after training the RBM, we compute n features (activations of the hidden units) for each of the s training samples. Thus, for each $k \in \{1, \ldots, n\}$ feature, we define a vector $\mathbf{f}_k = [f_r : r = 1, \ldots, s]$ that represents the values of feature k in all the training samples. Then, in the case of multisequence MRI data, which is typically used in many clinical scenarios such as the ones presented here, we measure MI between each feature $\mathbf{f}_k$ and the intensities of each c MRI sequence ($\mathbf{i}_c = [i_r]$) to quantify the statistical dependence between features and each sequence. Finally, for each feature, we combine the MI measure between $\mathbf{f}_k$ and each MRI sequence, as

$$MI_k(\mathbf{f}_k, \mathbf{i}_c) = \sum_c H(\mathbf{f}_k) + H(\mathbf{i}_c) - H(\mathbf{f}_k, \mathbf{i}_c), \tag{5}$$

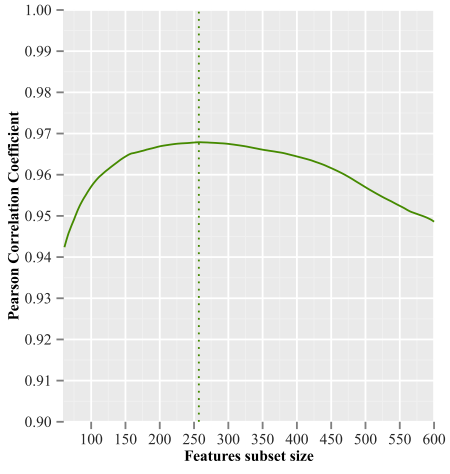
where H corresponds to the Shannon entropy. We will refer to MI_k as RBM-MI in order to express that the features are calculated by the RBM.

In RF, the contribution of each feature to decrease the impurity of training samples as they traverse the RF nodes can be evaluated with the MDI metric (Louppe et al., 2013). Although this estimate may be biased, it is still recommended, as obtaining unbiased feature importance estimations from tree-based ensemble methods is quite impractical (Louppe et al., 2013). MDI is computed using the Information Gain as splitting criteria in the nodes, which is equivalent to measure MI between the decision in the nodes and the class of the samples (Nowozin, 2012). We will denote this second component as RF-MDI.

The key idea in our proposal is to unify RBM-MI and RF-MDI into a common metric, in order to evaluate the overarching mapping between data and labels. Hence, for feature selection, we link MI measures between features and data (through RBM-MI) with MI measures between features and classes (through RF-MDI). From experiments, we observe that when we plot the RF-MDI and RBM-MI for feature k in descending order (Fig. 3, up) the curves are similar in shape, with a steep initial decrease. Then, features are gradually less important. We want to find the point that corresponds to the transition between important and unimportant features, both in terms of data representation and class label characterization. To this end, we measure the Pearson correlation coefficient between the sorted RF-MDI and RBM-MI measures for increasingly larger feature subsets; we note that the Pearson correlation coefficient is computed over the RF-MDI and RBM-MI measures, not the features themselves. Finally, the maximum in the Pearson correlation coefficient curve indicates the transition point γ between important and unimportant features (Fig. 3(b)). The final subset of selected features corresponds to the union of the best γ features sorted ac-



(a)



(b)

Fig. 3. Feature selection is based on RF-MDI and RBM-MI. a) RBM-MI (red) and RF-MDI (blue) of each feature are plotted in descending order. b) Pearson correlation coefficient between accumulating subsets of features MDI and MI. The dotted green vertical line marks the maximum of the Pearson correlation coefficient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Algorithm 1 Joint RBM-RF Mutual Information feature selection.

function FeatureSelection (**mi**, **mdi**)

Input: The vector sorted in descending order of RBM-MI values between each feature and the MRI sequences **mi**; the vector sorted in descending order of RF-MDI **mdi**

Output: The subset $\mathcal{S}' \subseteq \mathcal{S}$ containing the index of selected features

$\mathbf{pc} \leftarrow initialize\_zeros()$
for $j \in \mathcal{S}$ **do**
  $\mathbf{pc}_j \leftarrow pearson\_correlation(\mathbf{mi}_{0,\dots,j}, \mathbf{mdi}_{0,\dots,j})$
end for
$\gamma \leftarrow argmax(\mathbf{pc})$
$\mathcal{S}' \leftarrow get\_features(\mathcal{S}, \mathbf{mi}, \gamma) \cup get\_features(\mathcal{S}, \mathbf{mdi}, \gamma)$
return $\mathcal{S}'$

cording to RF-MDI and RBM-MI (c.f. Fig. 3 and Algorithm 1). As noted in (Ganz et al., 2015), using intersection instead of union can result in an empty set, although it can be prevented by including domain-specific knowledge. The proposed approach has the advantage of not requiring any pre-defined threshold (e.g. on the number of features). Still, if desired, one can define a minimum percentage or number of features to be evaluated through MI and MDI, and retrieve the union of those subsets.

3.2. Interpretability system

The interpretability of a machine learning model can be defined as the *human reasoning* on *how* a model captures the input data, and *why* a certain decision is made. In contrast to previous works, we acknowledge the relevance of both global and local interpretability as providers of complementary information.

Global interpretability is defined as the reasoning and identification of relevant inputs for the machine learning system as a whole. Although less model agnostic, it may help us answering *how* a model captures the (training) input. In the context of medical imaging, it is a description of what the system learns from a *population*. This may be used as a sanity check before deployment of the given machine learning model, since the expert can infer whether the way the model is capturing the input data is coherent with his prior knowledge on the problem. In addition, it is a valuable tool to unveil possible biases in the datasets introduced by the population subjects selection or data processing.

We define local interpretability as the reasoning and identification of the relevant inputs for a given decision, on a per-sample basis. Thus, locality is more related to *why* a decision is made, which becomes more relevant after deployment, especially to study failures. In the context of medical imaging, it can retrospectively help in understanding model decisions for a particular *subject*. Since our application is on image segmentation, local interpretability is at the voxel level and thus also patient-level.

Our approach is post-hoc in the sense that interpretation comes at a later stage to model training, instead of being embedded in the system itself. In this way, we avoid sacrificing complexity/performance of the model in favor of interpretability. In the following subsections we describe how global and local interpretability are implemented in the proposed approach.

3.2.1. Global interpretability

We propose to study the relationship between the input data (in our example applications, the different MRI sequences) and the RBM features. We define first a set containing all available n feature indices, i.e. $\mathcal{S} = \{1, 2, \dots, n\}$. Based on our feature selection method presented in Section 3.1.3, we can generate a reduced feature set $\mathcal{S}' \subseteq \mathcal{S}$, which is more suitable for model interpretation than the complete set of features. Some visible units may be grouped into z subsets of meaningful groups $\mathcal{G}_y$, such that $\mathcal{G} = \{\mathcal{G}_y : y = 1, 2, \dots, z\}$. Each $\mathcal{G}_y$ contains the indices of the visible units belonging to that meaningful group. For instance, when using image patches from multi-sequence MRI acquisitions we can group visible units belonging to the same sequence. In other words, the z-available image sequences define the groups $\mathcal{G}_y$ (see Fig. 2). In the extreme case where no meaningful groups can be defined, each visible unit v_i is a subset in itself. For interpretation, we compute the squared L2-norm of the weights connecting a hidden unit to the visible units of each group $\mathcal{G}_y$. This way, we determine the contribution of each group $\mathcal{G}_y$ to a hidden unit (i.e. feature). We repeat this procedure for all hidden units. We compute the squared L2-norm because a negative weight may still make a visible unit contribute positively to the hidden unit response, since the inputs are normalized with zero mean and unit variance. Additionally, it decreases the contribution of very small valued weights that would contribute to noise, which may impair the interpretability through visualization techniques. Algorithm 2 presents the complete procedure. Taking brain tumor segmentation as example, we identify the relevance of the different MRI sequences for a specific task (e.g.

Algorithm 2 Squared L2-norm computation for global interpretability.

function GlobalInterpretability ($\mathbf{W}$, $\mathcal{S}'$, $\mathcal{G}$)
Input: The weight matrix $\mathbf{W}$, the set $\mathcal{S}' \subseteq \mathcal{S}$ containing the index of selected features (=output of algorithm 1), and the set of meaningful groups $\mathcal{G}$
Output: Matrix with shape $[|\mathcal{G}|, |\mathcal{S}'|]$ of squared L2-norms $\mathbf{L} = [l_{y,j'}]$

```
for j' ∈ S' do
    for G_y ∈ G do
        l_{y,j'} ← Σ_{i'∈G_y} w_{i',j'}^2
    end for
end for
return L
```

Algorithm 3 Local test case specific feature selection.

function LocalFeatSel (x, $f(\cdot)$, n_{feat}, n_{x_p}, $g(\cdot), \mathcal{S}'$)
Input: A test sample x, a model $f(\cdot)$, the number of desired features n_{feat}, the number n_{x_p} of neighbors of x, and feature selection method $g(.)$
Output: Indices of the selected features $\mathcal{S}'' \subseteq \mathcal{S}'$

```
{x_p, d_{x_p}} ← Perturb(x, n_{x_p})
y_p ← f(x_p)
S'' ← g(x_p, y_p, d_{x_p}, n_feat)
return S''
```

Algorithm 4 Spatial feature relevance for local interpretability.

function SpatialFeatureRelevance ($\mathcal{I}$, $\mathcal{S}''$, $\mathcal{C}$, $\mathbf{W}$)
Input: the image space $\mathcal{I}$, the selected features for each test sample $\mathcal{S}''$ (=output of algorithm 3), the set of (pre-aligned) MRI sequences $\mathcal{C}$ (e.g. $\mathcal{C} = \{T_1, T_{1c}, T_2, FLAIR\}$), and the weights matrix $\mathbf{W}$
Output: Images with the voxel-wise importance for each sequence $\mathcal{H} = \{\mathbf{H}_c\}$

```
H ← initialize_zeros(I)
W ← normalize_weights(W)
for x ∈ I do
    p ← get_patch_indices(x)
    for j'' ∈ S'' do
        for c ∈ C do
            w_c ← sequence_weights(W, j'', c)
            for e ∈ p do
                H_c(e) ← H_c(e) + w_c(e)
            end for
        end for
    end for
end for
return H
```

segmentation of the complete tumor vs. normal tissues) by studying the weights connecting the hidden units of the most and least important features.

To facilitate interpretability, we sort the selected features in descending order of their importance (we took the RF-MDI as a measure of feature importance for the classifier itself). Then, we plot the squared L2-norm of the weights connecting the hidden units of the RBM to the subgroups of meaningful features (e.g. Fig. 4).

3.2.2. Local interpretability

We start by selecting the most meaningful features of a sample x (e.g. voxel of interest) by examining its neighborhood. By randomly perturbing the feature vector, which is the activations of the hidden layer of the RBM, of the selected sample x, a set of π synthetically generated samples ($\mathcal{X} = \{x_p : p = 1, \ldots, \pi\}$), with feature vector $\mathbf{f}_{x_p}$, is created. These synthetic neighbors are close to the original sample in the feature vector space. Then, a classification (y_p) for each neighbor is calculated with the classifier under analysis $f(\cdot)$ (RF in our machine learning system), as $y_p = f(\mathbf{f}_{x_p})$. Afterwards, a Ridge regression model ($g(\cdot)$) is trained on the synthetic set, using the output provided by the classifier under analysis as target, to select a pre-defined number of explaining features, corresponding to those yielding the strongest responses to the input, i.e. highest value of the product between the weights of the Ridge regressor and features. The output of $g(\cdot)$ is a subset $\mathcal{S}''$ of the available feature indices contained in $\mathcal{S}'$. Neighbors are weighted according to their euclidean distance to the original test case (d_{x_p}). Ridge regression has the advantage of being simple and extremely efficient, which allows it to be applied voxelwise.[2] The procedure is depicted in Algorithm 3. Here, we consider $\mathcal{S}'$ provided by a previous feature selection procedure, which selects features that highly correlate both with labels and the data. However, there is nothing that prevents Algorithm 3 from being applied while taking the full set of features into consideration. Additionally, this procedure can be used both in binary and multi-class problems.

Having the selected features, we could generate L2-norm plots similar to the ones proposed for global interpretability for each test sample. However, since we are dealing with images, it is more insightful to observe which parts of an image are more important for a given task, such as segmenting a particular tumor compartment. After selecting the features that better explain a prediction, we assume that all of those features must be equally taken into account for interpretation. However, the weights may be in different ranges. So, we independently normalize the absolute value of weights connecting each hidden unit to the visible units to $[0, 1]$. In our case, we predict the class of each voxel x based on the features computed by the hidden units of the RBM, which are extracted from a patch $\mathbf{p}$ centered on that voxel. Thus, we proceed by summing the weights connecting each selected hidden unit to the visible units of each sequence for all corresponding voxels contained in the patch centered on voxel x. By repeating this for all voxels in the image space $\mathcal{I}$, we obtain, for each MR sequence $c \in \mathcal{C}$, a corresponding image $\mathbf{H}_c$ that contains voxel-wise importance values. We denote this procedure as spatial feature relevance for local interpretability. The procedure is described in Algorithm 4.

The selection of locally relevant features is inspired by Ribeiro et al. (2016b) and motivated by some features being important for some classes, while other features may be important for some other classes. Thus, even though we selected a subset of important features before, some of them are more relevant depending on the sample under analysis. Contrasting with Ribeiro et al. (2016b), where a simpler, yet interpretable, model (Ridge regression in our case) serves as feature selector and explainer, we go further as to explain which parts of the input data mostly contributed to the features response of the sample under analysis. We realized this by projecting the MRI sequence relevance for a given feature back to the image space of the respective MR sequences. Additionally, we previously selected a subset of features that correlate both with the labels and the data, which differs from Ribeiro et al. (2016b).

[2] Instead of the proposed approach with Ridge, we could use stability selection with Lasso (Nicolai Meinshausen, 2010), which can select the optimal number of features. However, it still needs a threshold above which a feature is considered relevant. Additionally, it works by training several models on randomly chosen subsets of the training set, making it much more computationally demanding than Ridge.

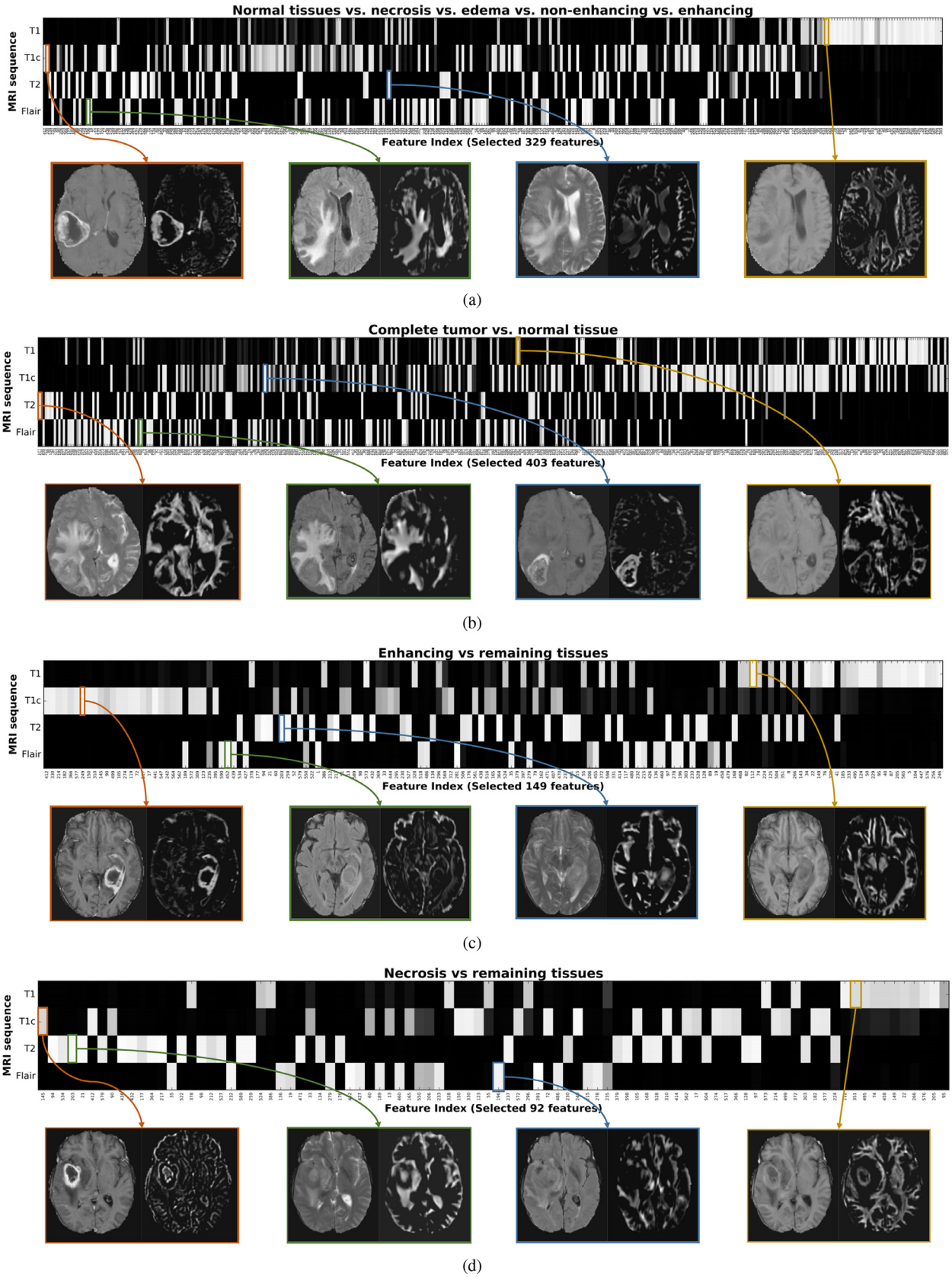


Fig. 4. Global interpretability on the BRATS model. Several segmentation tasks are studied: a) all tissues at once (multi-label), b) complete tumor vs. normal tissues, c) enhancing tumor vs. remaining tissues, and d) necrosis vs. remaining tissues. For each task we show: top) Squared L2-norm plots. Features are sorted from most to least important (left to right). Brighter means higher squared L2-norm of the weights connecting the hidden unit of a given feature to a given MRI sequence. Bottom) examples of pairs of MRI sequences (left) and feature maps (right).

4. Experimental setup

4.1. Databases

The proposed methodologies were applied to two segmentation problems with multisequence MRI data: Brain tumor segmentation and penumbra estimation in acute ischemic stroke, for which model predictions can be interpreted in the context of clinical expert knowledge and manual segmentation protocols (Menze et al., 2015; Maier et al., 2017).

4.1.1. Brain tumor segmentation

For this problem, we used the publicly-available BRATS 2013 database (Menze et al., 2015) of the MICCAI Brain Tumor Segmentation (BRATS) challenge. The database has three sets with different number of subjects: Training (30), Leaderboard (25), and Challenge (10). The Training set contains manual ground truth segmentations, distinguishing four tumor tissues: Necrosis, edema, contrast-enhancing tumor, and non-enhancing tumor. The evaluation of the Leaderboard and Challenge segmentation was performed via the online platform SMIR[3] for three tumor regions: Complete (all tumor tissues combined), core (necrosis + enhanced + non-enhanced), and enhancing tumor. For each subject there are four MRI sequences available with interpolated isotropic resolution of 1 mm: T1-weighted (T1), gadolinium-enhanced T1 (T1c), T2-weighted (T2), and Fluid-attenuated Inversion Recovery (FLAIR). All sequences are already rigidly aligned, and skull-stripped. Further pre-processing included bias field correction (Tustison et al., 2010), and normalization of the intensities in each MRI sequence with a histogram standardization method (Nyúl et al., 2000). Finally, we normalized the intensities of brain voxels to zero mean and unit variance.

We chose the BRATS 2013 database due to two reasons: First, the ground truth data are manual segmentations obtained by the fusion of four expert raters. These expert raters followed a manual segmentation protocol (Menze et al., 2015) to acquire the ground truth data. Hence, it enables us to interpret our machine learning system with respect to this protocol. Second, the dataset contains preoperative brain tumor images only. In contrast to postoperative images, treatment-related imaging changes (e.g. radiation necrosis Mullins et al., 2005) are absent in preoperative images thus rendering an evaluation of our interpretation methodologies less complicated.

4.1.2. Penumbra estimation in acute stroke

For investigating the penumbra estimation in acute ischemic stroke, we employed the Stroke Perfusion Estimation (SPES) database of the MICCAI Ischemic Stroke Lesion Segmentation (ISLES) Challenge (Maier et al., 2017). The Training dataset includes 30 subjects with publicly available manual ground truth segmentations, while the Challenge set is composed of 20 subjects. As in BRATS, the results for the Challenge set are computed by an online platform.[4] Seven MRI sequences, comprising structural and physiological sequences, were available: T1c, T2, Diffusion Weighted Imaging (DWI), cerebral blood flow (CBF), cerebral blood volume (CBV), time-to-peak (TTP), and time-to-max (Tmax). All sequences are already rigidly registered to the T1c sequence with image resolution of 2 mm and skullstripped. Further pre-processing included the bias field correction (Tustison et al., 2010), and normalization of the intensities with a histogram standardization method (Nyúl et al., 2000) for the T1c, T2, and DWI sequences. Additionally, we clipped Tmax intensity values above 60 (Tmax $>$ 6s threshold, as followed by the manual segmentation protocol experts Maier et al., 2017). Finally, we normalized the intensities of brain voxels to zero mean and unit variance.

Table 1
Hyperparameters of the RBM and RF of our machine learning system. In RF, when not indicated, default values were used.

Database	Algorithm	Hyperparameter	Value
BRATS SPES	RBM	Hidden units	600
		Mini-batch	32
		Gibbs samp. steps	1
		W init.	$\mathcal{N}(0,\ 1\times10^{-4})$
		a, **b** init.	0
	RF	Trees	200
		Split. crit.	Info. gain
BRATS	RBM	Patch size	$9\times9\times9$
		Epochs	262
		Initial ϵ; Final ϵ	1×10^{-4}; 4×10^{-7}
		Initial η; Final η	0; 0.5
		L1; L2	1×10^{-3}; 2×10^{-2}
SPES	RBM	Patch size	$5\times5\times5$
		Epochs	498
		Initial ϵ; Final ϵ	1×10^{-3}; 1×10^{-5}
		L1; L2	2×10^{-4}; 2×10^{-4}

4.2. Model training & parameters

Around 40,000 samples were extracted from each subject, and classes were balanced by having 50% of normal tissue and 50% of total lesion tissue (in BRATS we further approximately balance the sampling of tumor tissues). The RBM consisted of 600 hidden units (=features). For training the RBM, the learning rate (ϵ) was kept constant for the first 10 epochs, and then linearly decreased until the end of the training. In the case of momentum (η), it was kept constant until epoch 100, and then linearly increased until the end of training; no momentum was used in SPES. For penumbra estimation we thresholded the probabilistic output of the RF at 0.6 (empirically found in the validation set). The remaining hyperparameters of the RBM and RF are shown in Table 1. For training the RBM, we extracted the patches centered on a given voxel from all MRI sequences. Then, the patches are represented as a 1D vector and fed into the visible layer of the RBM. So, for brain tumor segmentation the visible layer has $4\times9\times9\times9=2916$ units, while for penumbra estimation it is $7\times5\times5\times5=875$ (see Fig. 2). For the local interpretability method, each image voxel's features were perturbed to generate 2400 synthetic samples, and the 10 most representative features were selected. We set the regularization parameter λ of the Ridge regression model to 1.0. We used the RF implementation in Scikit-learn (Pedregosa et al., 2011); the hyperparameters that are not defined in Table 1 were set to default values. We used LIME[5] for generating the neighborhood of the points and select the local relevant features for local interpretability. The implementation of the proposed algorithms is available online.[6]

In this paper, we focus on enhancing the interpretability of the machine learning system at hand, thus a thorough performance evaluation is out of scope. Nevertheless, evaluating the segmentation to some extent is imperative to assess if the model is learning. Thus, we report the Dice Similarity Coefficient (Dice) for BRATS and SPES, as well as the Average Symmetric Surface Distance (ASSD) for the latter, as defined in (Menze et al., 2015; Maier et al., 2017).

[3] www.smir.ch/BRATS/Start2013.
[4] www.smir.ch/ISLES/Start2015.
[5] github.com/marcotcr/lime .
[6] github.com/sergiormpereira/EIML .

5. Results

5.1. Feature selection

In Table 2, we present segmentation results using features selected with the proposed method in BRATS and SPES. In order to have our approach compared with other methods for automatic feature selection, we also present results when features are selected by Embedded or Wrapper feature selection-based approaches. Embedded methods are represented by Lasso, Elastic Net, and stability selection with Lasso (Nicolai Meinshausen, 2010). Recurrent feature elimination using a linear kernel Support Vector Machine (RFE L-SVM), a RF, or a Ridge are categorized as wrapper approaches. We note that the results reported in Table 2 were obtained using the RF classifier, but, with features selected by those feature selection approaches.

In BRATS, with the proposed approach, we obtained a set of 329 features. The remaining methods selected the following number of features: elastic net – 440, Lasso – 168, RFE L-SVM – 420, RFE RF – 350, RFE Ridge – 400, and Lasso stability selection –572. For the case of SPES challenge data, the proposed approach yielded a subset of 117 features. In this application, the other methods selected the following number of features: Elastic net – 383, Lasso – 376, RFE L-SVM – 200, RFE RF – 350, RFE Ridge – 550, and Lasso stability selection –583. See # features BRATS/SPES in Table 2. It is important to note that these embedded and wrapper methods were executed in a cross-validation scheme to find the optimal parameters and features. The only statistical difference was observed when comparing the proposed method with RFE RF in SPES, under a paired Wilcoxon Signed-Rank Test with significance level of $\frac{0.05}{7}$ (Bonferroni-correction).

5.2. Comparison with other segmentation methods

Although the focus of this work is on interpretability of our machine learning system instead of performance, we compared our results with the contestants of the on-site BRATS 2013 (Table 3) and SPES 2015 (Table 4) challenges. The compared methods include proposals based on ensembles of randomized trees, such as Festa, Meier, Reza, and Tustison, in BRATS; or CH-Insel, DZ-Uzl, and BE-Kul2, in SPES. The method CA-Usher, in SPES, is built over a supervised Representation Learning algorithm (CNN).

We observed that in BRATS the set of important features changes accordingly to the task at hand (c.f. Section 5.3.1). For example, when we segment the complete tumor as a binary problem, or all tissues as a multi-label segmentation problem. Hence, motivated by this observation, and inspired by Meier et al. (2014b), we evaluated a hierarchical approach: First, we segmented the complete tumor, then we segmented the tumor tissues inside the previously defined region of interest. In Table 3 it is possible to observe that the detection of the complete tumor improved with the hierarchical approach, suggesting that different features are useful for different tasks. In BRATS (Table 3), the proposed model achieved a lower Dice compared to Tustison. However, it is on pair with the other top methods. In SPES (Table 4), the obtained results are comparable with the algorithms in the mid-table positions.

5.3. Interpretability

We present two case studies for interpretability: brain tumor segmentation (Section 5.3.1) and penumbra estimation in acute ischemic stroke (Section 5.3.2). In both cases, we present global and local interpretation results.

Table 2
Comparison of feature selection methods on BRATS and SPES data. The percentages indicate the fraction of features retained after feature selection. The metrics in the right correspond to the Dice on BRATS 2013 (Leaderboard and Challenge) and SPES.

				BRATS Leaderboard			BRATS Challenge			SPES
Selection approach	Method	# features BRATS	# features SPES	Complete	Core	Enh.	Complete	Core	Enh.	Penumbra
Embedded	Elastic Net	440 (73%)	383 (64%)	0.73 ± 0.22	0.60 ± 0.28	0.58 ± 0.33	0.81 ± 0.05	0.75 ± 0.14	0.72 ± 0.10	0.75 ± 0.14
	Lasso	168 (28%)	376 (63%)	0.73 ± 0.22	0.57 ± 0.28	0.56 ± 0.32	0.81 ± 0.05	0.73 ± 0.14	0.71 ± 0.10	0.74 ± 0.14
	Lasso stability selection	572 (95%)	583 (97%)	0.74 ± 0.21	0.61 ± 0.28	0.57 ± 0.33	0.80 ± 0.05	0.75 ± 0.14	0.72 ± 0.10	0.74 ± 0.14
Wrapper	RFE L-SVM	420 (70%)	360 (60%)	0.74 ± 0.22	0.61 ± 0.28	0.58 ± 0.33	0.80 ± 0.05	0.74 ± 0.14	0.72 ± 0.10	0.75 ± 0.14
	RFE RF	350 (58%)	200 (33%)	0.74 ± 0.21	0.60 ± 0.28	0.58 ± 0.33	0.80 ± 0.05	0.74 ± 0.14	0.72 ± 0.10	0.75 ± 0.14
	RFE Ridge	400 (67%)	550 (92%)	0.74 ± 0.21	0.60 ± 0.28	0.57 ± 0.33	0.80 ± 0.05	0.74 ± 0.13	0.72 ± 0.10	0.74 ± 0.14
	Proposed – All feat.	600 (100%)	600 (100%)	0.74 ± 0.21	0.61 ± 0.28	0.58 ± 0.33	0.81 ± 0.05	0.74 ± 0.13	0.72 ± 0.10	0.75 ± 0.14
	Proposed – Sel. feat.	329 (55%)	117 (20%)	0.73 ± 0.22	0.60 ± 0.28	0.58 ± 0.33	0.81 ± 0.04	0.74 ± 0.13	0.72 ± 0.09	0.74 ± 0.14

Table 3
Comparison with other methods on BRATS 2013 challenge set. Results obtained from (Menze et al., 2015).

Method	Dice		
	Complete	Core	Enh.
Cordier	0.84	0.68	0.65
Doyle	0.71	0.46	0.52
Festa	0.72	0.66	0.67
Meier	0.82	0.73	0.69
Reza	0.83	0.72	0.72
Tustison	0.87	0.78	0.74
Zhao	0.84	0.70	0.65
Proposed – All feat.	0.81	0.74	0.72
Proposed – Sel. feat.	0.81	0.74	0.72
Proposed – Hierarchical	0.84	0.74	0.71

Table 4
Comparison with other methods on SPES challenge set. Results obtained from (Maier et al., 2017).

Selection method	Penumbra	
	Dice	ASSD
CH-Insel	0.82 ± 0.08	1.65 ± 1.40
DE-Uzl	0.81 ± 0.09	1.36 ± 0.74
BE-Kul2	0.78 ± 0.09	2.77 ± 3.27
CN-Neu	0.76 ± 0.09	2.29 ± 1.76
DE-UKF	0.73 ± 0.13	2.44 ± 1.93
BE-Kul1	0.67 ± 0.24	4.00 ± 3.39
CA-Usher	0.54 ± 0.26	5.53 ± 7.59
Proposed – All feat.	0.75 ± 0.14	2.43 ± 1.93
Proposed – Sel. Feat.	0.74 ± 0.14	2.48 ± 2.04

5.3.1. Brain tumor segmentation

Since the segmentation of brain tumors and their subcompartments reflect a multi-label classification problem, we can define different segmentation tasks: all tissues at once, complete tumor vs. normal tissue, enhancing tumor vs. remaining tissues, or necrosis vs. remaining tissues. The first task is a multi-label classification problem, where the target labels are all tissues – normal, necrosis, edema, non-enhancing, and enhancing tumor. The other tasks are binary classification problems; in the case of complete tumor vs. normal tissue, we fuse all tumor tissues of the manual segmentation into just one class, and we use it for training. These different segmentation tasks serve the purpose of interrogating the machine learning system at hand on the usefulness of features extracted from the different MRI sequences. To leverage interpretability, we selected important features with the proposed feature selection method (Section 3.1.3), leading to the following number of selected features: 329 (all tissues at once), 403 (complete tumor vs. normal tissue), 149 (enhancing tumor vs. remaining tissues), and 92 (necrosis vs. remaining tissues). For each resulting feature set, we trained a RF model and performed global and local model interpretation analyses.

5.3.1.1. Global interpretation. We can interpret the model from a global point of view by inspecting how it learned the input data. Fig. 4 shows the squared L2-norm plots for global interpretability, as well as some feature maps representative of each zone of importance alongside with the sequence to which they are more related to (in terms of mutual information). In Fig. 4 it is possible to observe that features encoding information from the T1 sequence are mostly relegated to the tail of the important features. In contrast, features computed by hidden nodes that were strongly connected to T1c, T2, or FLAIR are given more importance. In very specific tasks, such as segmenting enhancing tumor (Fig. 4), or necrosis (Fig. 4), some particular sequences are preferred, such as T1c, or T2, respectively. On the other hand, the top features of more complex (multi-label) tasks, such as segmenting all tissues at once, have a higher mixture of features strongly connected to T1c, T2, and Flair. A similar behavior is observed when we segment the complete tumor (Fig. 4), with the difference that T1c is less important, because FLAIR and T2 are sufficient to delineate the lesion as a whole. Interestingly, the hidden nodes of the RBM are more connected to one specific MRI sequence, instead of collecting information and combining multiple sequences. This is confirmed by the feature maps that can depict some specific tissues. For instance in Fig. 4, left, it is conspicuous for enhancing tumor, or in Fig. 4, second from left, the feature map appears to enhance edema.

5.3.1.2. Local interpretation. From the local interpretability point of view, we studied the local spatial feature relevance for assessing how the input data was used for voxel-wise predictions in a given test subject (Figs. 5 and 6). Similarly to the global interpretability, we studied the local interpretability for the different tumor segmentation tasks (i.e. “all tissues at once”, “complete vs. normal tissue”, “enhancing vs. remaining tissues”, and “necrosis vs. remaining tissues”). From Fig. 5 it is observed that the identification of each class is attributed to a subset of the available MRI sequences coherent with observations from the global analysis. For instance, enhancing tumor is strongly linked to T1c, while necrosis extracts more information from T2, but also from T1c in some extent. In the case of complete tumor (Fig. 6), FLAIR resembles to play an important role, although T1c contributes considerably in the region of enhancing tumor. In contrast to global interpretability, the local interpretability analysis allows us to better disentangle the relevance of the different sequences for a particular patient and image region as well as to study the cause of false positive segmentations. As an example, in Fig. 6 on the superior part of the brain there are some false positive tumor segmentations visible. We observed that they are more related to T1c and FLAIR than to the remaining sequences. As discussed below, the local interpretability analysis allowed us to find potential causes and pre-processing related issues that led to these false positives.

5.3.2. Penumbra estimation

Contrasting to the multi-class brain tumor segmentation, in SPES the task is binary and aims at segmenting the penumbra region. On the other hand, SPES contains seven MRI sequences including structural and physiological information, in contrast to the four structural MRI sequences in BRATS. In this dataset, the number of selected features by the proposed approach was 117.

5.3.2.1. Global interpretation. Fig. 7 shows the squared L2-norm plots for the global interpretability of the model, as well as some feature maps and the MRI sequence to which the respective hidden unit is most connected. First, we can observe that some specific MRI sequences contribute much more than others to the most relevant features. The first three most important features come from the TTP sequence, while overall the Tmax sequence has the largest number of most important features. Observing the feature maps in Fig. 7, they characterize the stroke region either as a hypointense or hyperintense area. MRI sequences such as DWI, T1c, and T2, have some features strongly connected to them for the most important features, but appear mainly on the least important section of the ranked features in Fig. 7 top. Interestingly, contrasting to BRATS (Fig. 4) where features are mostly related to just one MRI sequence, in penumbra estimation some features are computed from both the DWI and T2 sequences. Finally, the CBF and CBV sequences are barely represented.

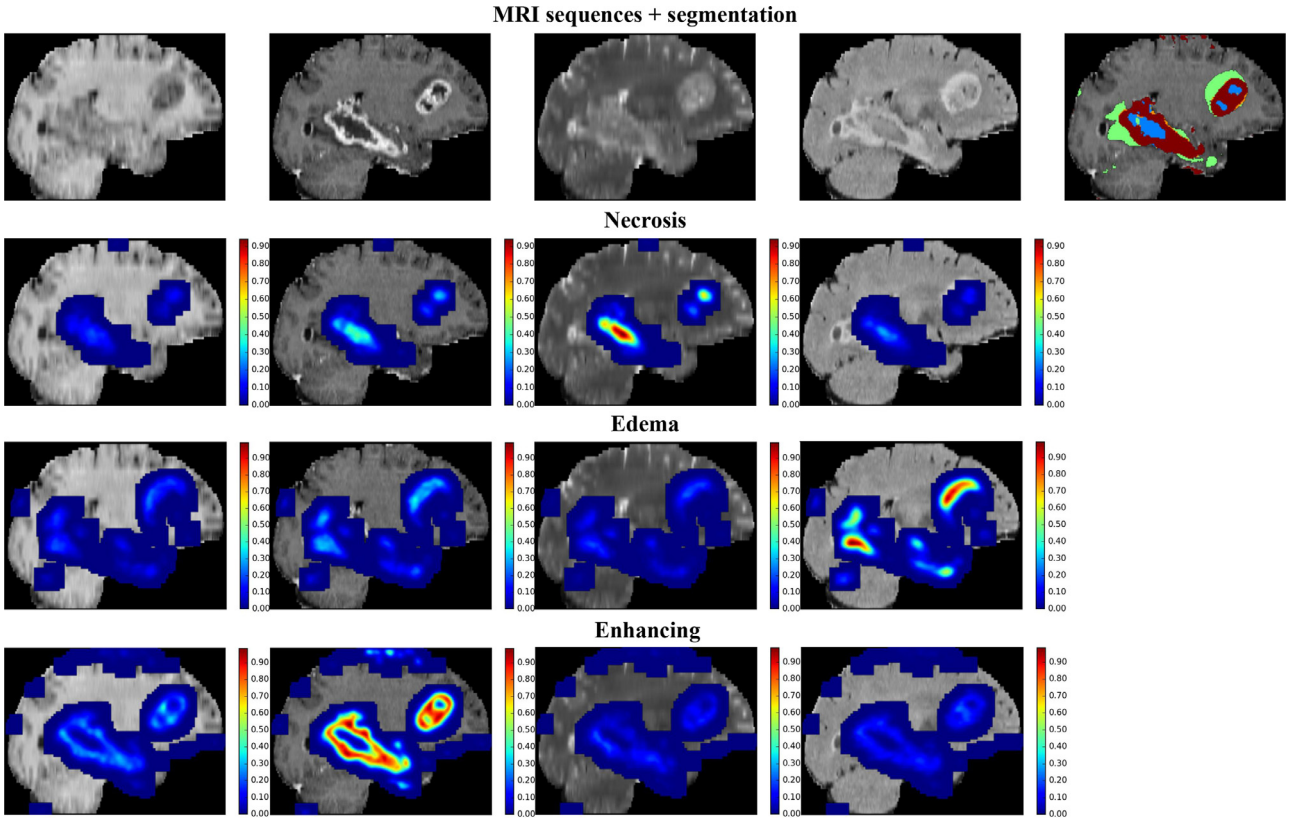


Fig. 5. Spatial feature relevance for local interpretability of the Challenge subject 0310 in BRATS for the task of segmenting all tissues at once (multi-label classification). From left to right, we show the T1, T1c, T2, and FLAIR sequences, as well as the obtained segmentation in the first row. In the segmentation, the tumor tissues are: blue – necrosis, green – edema, orange – non-enhancing, and red – enhancing tumor. Each row corresponds to how the input data was used for predicting each class. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.3.2.2. Local interpretation. Local spatial feature relevance for penumbra estimation is presented in Fig. 8. It is possible to observe that Tmax and TTP are the sequences from which the model takes more information. TTP appears with higher magnitude for relevance, but in the posterior part of the segmentation it is lower compared to Tmax. Features related to the other MRI sequences are less preferred than Tmax and TTP, with T1c appearing with a larger contribution to approximate the overall stroke region segmentation, and CBV appearing to be the least important for voxel-wise predictions.

5.3.2.3. Removal of MRI sequences. Given the observations from the global and local interpretations that CBV and CBF play a minor role in penumbra estimation, we experimented to train the system without those MRI sequences. The system trained without the CBV sequence results in ASSD of 2.58 and Dice of 0.74. When we further removed both the CBV and CBF sequences the results were: ASSD – 2.54 and Dice – 0.74. Comparing with Table 4, we can observe that the results are equivalent to those using all the available MRI sequences.

6. Discussion

Machine learning systems are pervasively being adopted as decision-support systems in critical fields, like the medical domain. At the same time, the models are increasingly complex for the sake of performance. This may pose a problem in their adoption due to trust reasons, as it is difficult to explain the model and/or the respective predictions. Thus, there is a need to understand *how* a model learned the input data, and *why* a certain prediction was made. In this paper, we propose a novel methodology to enhance the interpretability of automatically extracted machine learning features. We also investigated the notion of global and local interpretability. Global interpretability provides insights as to *how* the model learned the data from a population. This allows us to infer if the studied model is coherent with the experts knowledge. On the other hand, local interpretability leverages the understanding on *why* a prediction on a subject-specific level was made. A limitation of interpretability systems is that there is no quantitative metric available to measure interpretability of a machine learning system yet. Thus, we base our discussion on extensive comparison with medical domain knowledge. The proposed machine learning system encompasses a RBM as representation mapping stage and a RF as task-related learner. High-dimensional feature vectors may impair interpretability. Thus, we propose a MI-based feature selection scheme that simultaneously take into account the mapping between the input data and its representation (features), and from the representations to the task at hand (labels).

6.1. Joint RBM-RF approach for feature selection

From Table 2 we observed that all feature selection methods perform similarly, both in mean and standard deviation; the only statistical difference was found when comparing the proposed approach with RFE RF in SPES. However, just by looking at the metrics, some embedded wrapper methods seem to achieve slightly better performances. However, they were evaluated in a cross-validation scheme to find the best parameters and number of features. Moreover, Wrapper recursive feature elimination schemes require training several models with progressively smaller feature vectors. The embedded methods do not require the recursive fea-

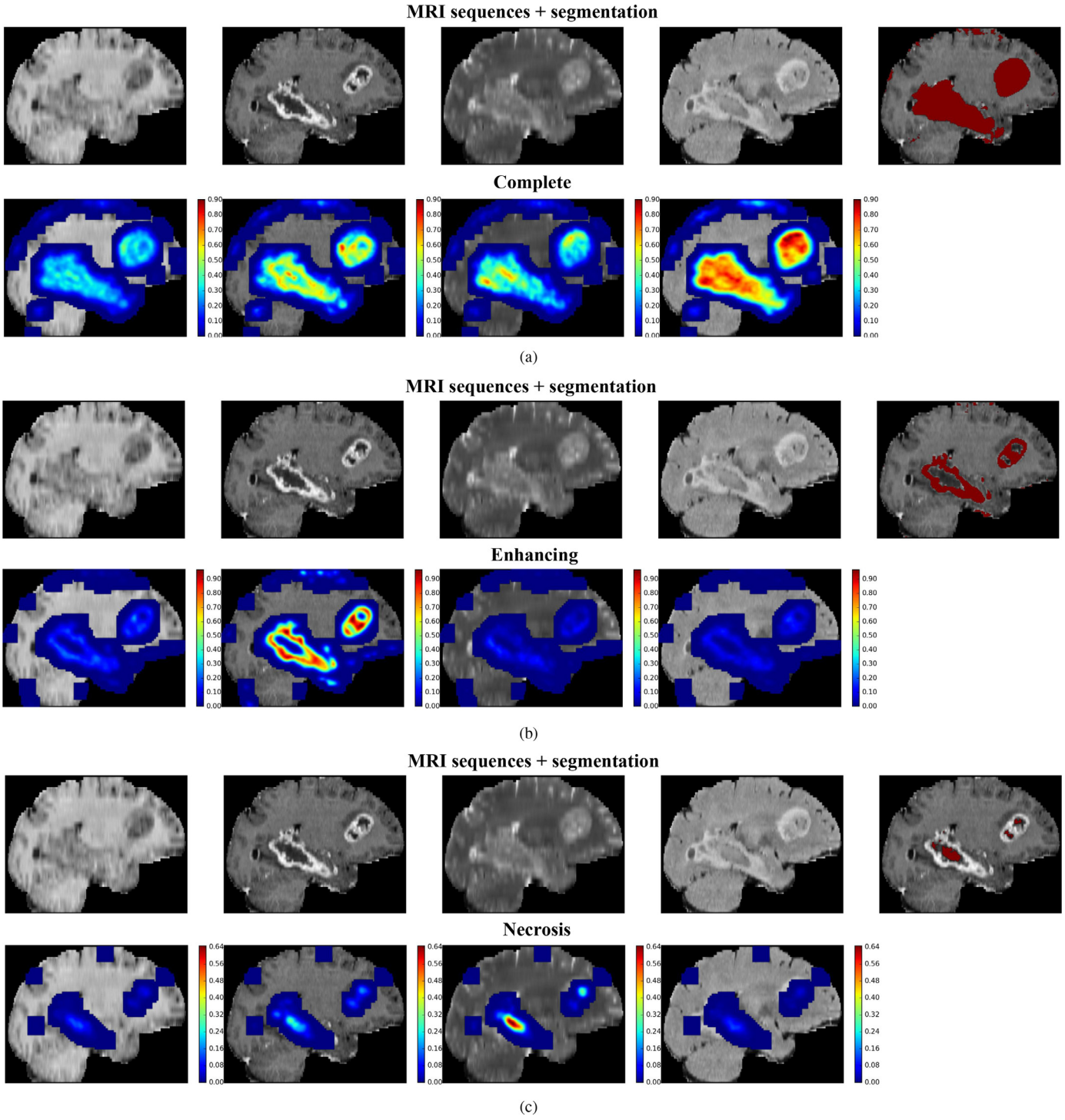


Fig. 6. Spatial feature relevance for local interpretability of the Challenge subject 0310 in BRATS for the tasks: a) Complete tumor vs. normal tissues, b) enhancing tumor vs. remaining tissues, and c) necrosis vs. remaining tissues. From left to right we show the T1, T1c, T2, and FLAIR sequences, as well as the obtained segmentation in the first row of each task.

ture elimination scheme. However, the selected features may be optimal for the used learner, but not for the model that we must use in the end. Contrarily, by combining RBM-MI and RF-MDI, the proposed approach offers the advantage of automatically selecting the optimal number of features and does not require a threshold to be defined, nor a recursive feature elimination scheme. The main motivation for feature selection in the context of this paper is to choose features that both correlate with the data and the labels, and to leverage interpretability. In that sense, the proposed approach provides more satisfactory results than the other methods, by decreasing the dimension of the feature vector to almost half in BRATS, and to around 20% in SPES (only Lasso in BRATS achieves a more compact feature subset). Moreover, the proposed feature selection approach does not impact the segmentation performance, when comparing results to a model using all 600 RBM-derived features. Thus, the hypothesis that selecting features with high mutual information both with labels and the data is viable. In this experiment, we proposed to use Pearson Correlation Coef-

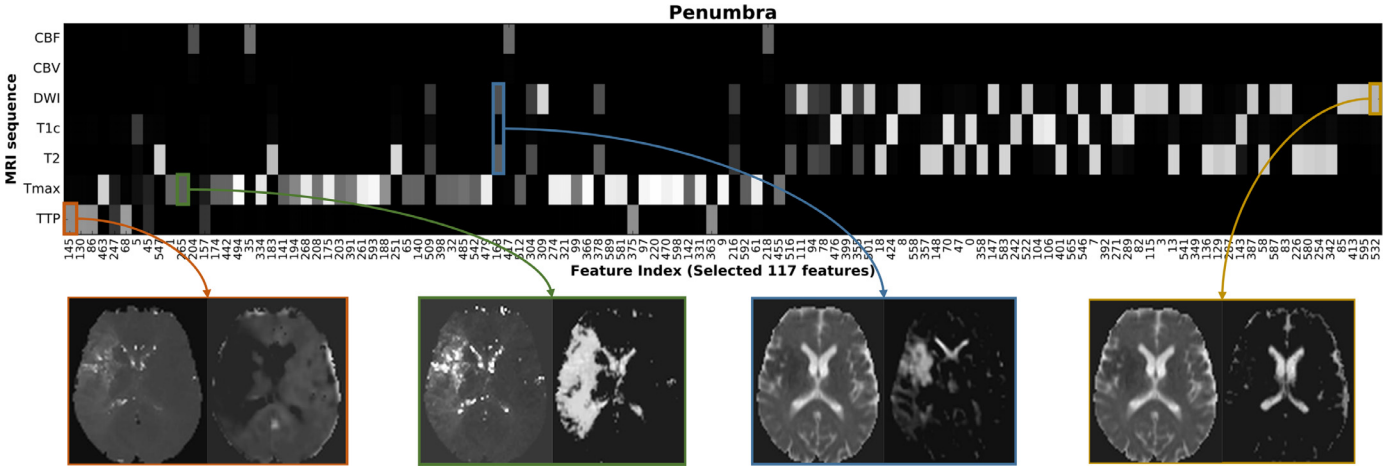


Fig. 7. Global interpretability on the SPES model. Top) squared L2-norm plots. Features are sorted from most to least important (left to right). Brighter means higher squared L2-norm of the weights connecting the hidden unit of a given feature to a given MRI sequence. Bottom) examples of pairs of MRI sequences (left) and feature maps (right).

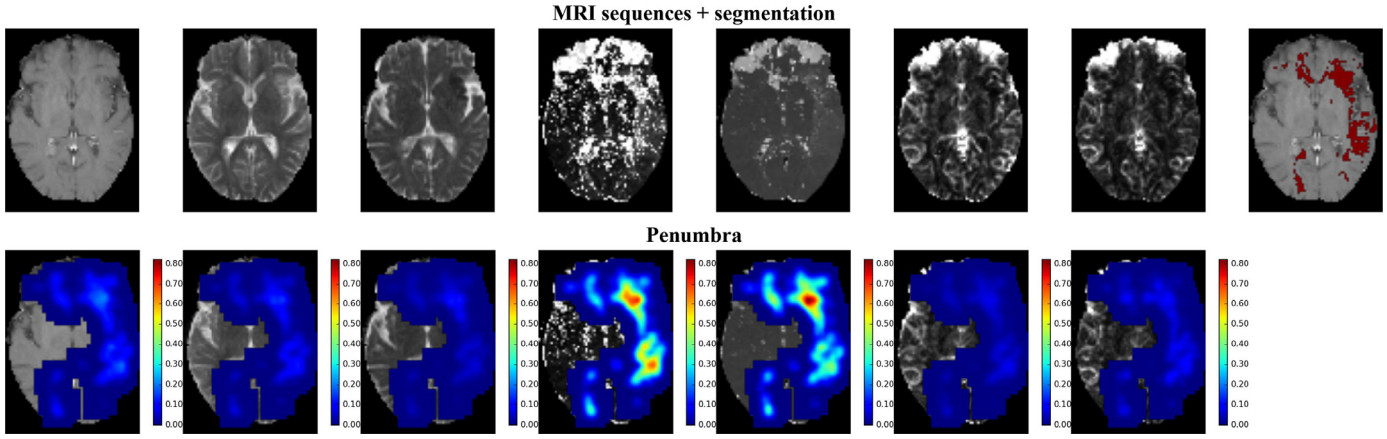


Fig. 8. Spatial local interpretability of the SPES Challenge subject 1. Top) MRI sequences and segmentation. From left to right we show the T1c, T2, DWI, Tmax, TTP, CBV, and CBF sequences, as well as the obtained segmentation. Bottom) local interpretability maps.

ficient computed over the RBM-MI and the RF-MDI measures to detect the transition between important and unimportant features. This choice came from the observation of the decreasing regime of both metrics, and its empirical nature represents the main limitation of the feature selection approach.

At the same time, this big reduction in the number of features imply that although RBM can automatically compute features, many of them can be useless in the presence of more powerful ones. This may be caused, in part, by the unsupervised nature of RBM, since it does not know for which task the features are going to be employed (Larochelle and Bengio, 2008). Although models using a high number of features can be prone to overfitting, in our experiments we did not observe such tendency, probably due to the robustness of RF (Criminisi and Shotton, 2013).

Observing that the subset of important features changes with the task at hand, we devised a hierarchical approach in BRATS. This allowed us to improve the segmentation of complete tumor. With this approach, we are on pair with the methods of the on-site results of BRATS 2013 Challenge, with the exception of the winner of that edition. We note that despite these methods being from 2013, they still remain as representative of RF-based approaches. All of them rely on hand-crafted features, while ours is based on an unsupervised representation mapping algorithm. In SPES, the proposed machine learning system is positioned on pair with the mid-table methods. However, top methods incorporate expert prior knowledge. CH-Insel includes atlases information, and presence of the voxel in the ipsi- or contralesional side. Additionally, both CH-Insel and De-Uzl compute features of symmetry in relation to the mid-sagittal plane. This kind of information cannot be captured by a representation learning algorithm. Even so, although the proposed system is based on an unsupervised model, it achieved better metrics than a CNN-based proposal (CA-Usher). We note, however, that the CA-Usher team achieved high performances in terms of metrics in the training set (similarly as all the other teams). This behavior may be related to overfitting on the training by this team, which may have been alleviated by the unsupervised nature of our approach.

6.2. Interpreting automatically extracted features in brain tumors

We can define several segmentation tasks in brain tumor image analysis. This allows us to interrogate and interpret the machine learning system and assess if it is learning well, according to clinical expert knowledge on the problem. Observing the squared L2-norm plots of the RBM weights connecting the hidden and visible units representing each MRI sequence in Fig. 4, we obtained insight into which sequences are more important for the different tasks. When we segment all tissues at once (Fig. 4), the most important features extract information from the T1c, T2, and FLAIR sequences. As expected, since T1 adds less information to the other ones, the features connecting strongly to this sequence are the least important, or appear sparsely represented in the most im-

portant features. From a clinical point of view, this result is valid for pre-operative brain tumor images as contained in the BRATS 2013 data set. In the feature maps, we note that specific patterns of enhancing tumor and edema were extracted from T1c and FLAIR. Fluid-filled compartments are mostly enhanced in T2, while T1 encodes the fatty tissue, mainly. This global interpretation is in line with clinical domain knowledge, and, hence, allows us to conclude that the model is correctly utilizing the imaging information. From the local interpretation results, in Fig. 5, we can observe not only which sequence contributes the most for each label class, but also study this contribution with respect to image regions. For example, segmentation of edema is mainly based on FLAIR. However, according to Menze et al. (2015), segmentation of edema is also based on T2. The reason for the model to prefer FLAIR may be because it can differentiate edema from the cerebrospinal fluid, which is contained e.g. in the ventricles. T1c is the second choice in predicting edema. Although one could expect T2 to appear after Flair, the enhancing rim reflects a strong prior on the extent of the tumor core (hence limiting the extent of edema, too), at least in the images of BRATS 2013, which are mostly high-grade glioma. Thus, since we deal with patches in our system, the model may learn that close to edema, features with high response on T1c may exist. As expected, T1c clearly dominates the predictions of enhancing tumor.

6.2.1. Interpreting the binary brain tumor segmentation tasks

Regarding the binary brain tumor segmentation tasks, we observed that among the most important features for complete tumor (Fig. 4) there are features strongly related to all sequences, contrasting, for instance, with segmentation of the enhancing tumor, or necrosis. This is due to the higher variability of the tumoral tissues included in the “complete tumor” region, than in the other binary tasks where the tissue type shows much lower intensity variability across sequences. For the binary task of segmenting the complete tumor as a whole, the T2 and FLAIR sequences are more important. This is in accordance with the manual segmentation protocol used in BRATS, where the complete tumor was firstly defined based in those same MRI sequences (Menze et al., 2015). From the first (i.e. most-left) feature map in Fig. 4, it can be observed that the hypointense portion of the T2 image is encoded in the respective feature map, including mainly areas of white matter and solid tumor tissue. Interestingly, some T1c-based features appear to capture intra-tumoral regions. This can be observed in the third (from left to right) feature map of Fig. 4, and in the local explanation of predictions in Fig. 6, where T1c is important in the enhancing rim region. This observation shows that the Ridge regression is able to identify locally important features, indeed. However, following results from both global and local analysis, features derived from FLAIR appear to be the most dominant for defining the complete tumor, in general. When we segmented the enhancing tumor against all the other tissues, the most important features are provided by hidden units with their weights strongly connected to the T1c visible units. In Fig. 6, we also observed that locally the T1c sequence completely dominates over the other ones. This was expected, since enhancing tumor region is characterized by the T1c image. Finally, in the necrosis segmentation task, features that are strongly connected to the T2 sequence appeared more often among the most important features in the global squared L2-norm plots (Fig. 4). However, some T1c-based are also ranked among the top. For the patient case 0310, shown in Fig. 6, we can see two lesions, each with necrotic tissue. Interestingly, for segmenting the larger necrotic core of the more central lesion, the T2 sequence seems to be more relevant than T1c. In contrast, for segmenting necrosis in the smaller lesion both sequences appear equally important. This observed complementarity of T1c and T2 is in line with the manual segmentation protocol (Menze et al., 2015), since both T1c and T2 are used for defining the tumor core and for differentiating the non-enhancing part of the tumor from necrosis and enhancing tumor. Necrotic regions are mostly surrounded by the enhancing tumor, hence the importance of the T1c for this task. Indeed, in the feature map related to this sequence, shown in Fig. 4, the enhancing rim appears completely dark, while the inner part, corresponding to necrosis, is enhanced. In this way, we are able to verify and conclude that the system learned the correct relations in the data. We emphasize, however, that these observations apply for pre-operative acquisitions only. In a post-operative setting the relevance of the different MRI sequences for tumor segmentation is different (Meier et al., 2017).

6.2.2. Further considerations

Apart from studying the relevance of the different sequences for tumor tissue predictions, we observed for patient case 0310 in Fig. 6 some misclassified tumor regions on the top of the brain. The dominant source for this misclassification are features related to the T1c or FLAIR sequences. Hence, these misclassifications are probably caused by errors of the skull-stripping procedure, which was insufficient for this image region (remaining extra-cerebral tissue such as e.g. meningeal tissue that typically appears enhanced in T1c, similarly to tumor tissue).

All the previously mentioned observations suggest that, despite being a completely unsupervised algorithm, the RBM is able to identify tissue patterns and the most important spatial features of the imaged pathologies at hand. Moreover, by inspecting the strength of the weights we can identify which inputs were considered more important for each feature. Taking these results into account, it is clear that RBM computes representations that are more important depending on the task at hand. This is confirmed by our hierarchical approach yielding improved results over the “segment all at once” scheme, and more similar to those obtained by handcrafted features.

Although some sequences are clearly less important than others, we observed a degradation in performance if some MRI sequence is removed. This is in accordance with Havaei et al. (2016), since the authors observed a performance drop when some sequence is missing, too. Brain tumors are characterized by being a heterogeneous kind of lesion, with some portions conspicuous only in some MRI sequences. As we can observe in Figs. 4–6, all the four MRI sequences have some importance for some task.

6.3. Interpreting automatically extracted features in acute ischemic stroke

We also evaluated the proposed methodologies in a penumbra estimation problem, using the SPES database of the ISLES2015 MICCAI challenge. During acute ischemic stroke, the most severely affected region, the core lesion, consists of irreversibly damaged tissue. Penumbra refers to the larger region of dysfunctional, but salvageable, tissue at risk of infarction. Hence, it is imperative to treat this region as fast as possible. The core lesion is conspicuous in DWI, but penumbra is not (Copen et al., 2011; Straka et al., 2010). Additionally, the core lesion is smaller than the penumbra region. Possibly because of this, features computed by hidden nodes that are strongly connected to the DWI sequence were not ranked among the most important features in the squared L2-norm plot for global interpretability, Fig. 7. If the system heavily relied on DWI, it could underestimate the penumbra extension (observe third feature map in Fig. 7, which relies both on DWI and T2). Conversely, the penumbra is better visualized in Perfusion Weighting Imaging. CBV corresponds to the blood volume in the area. Regions of low CBV correlate with the core and final outcome of the infarction; however, we are interested in predicting the complete penumbra. In turn, CBF is related to the supply of oxygen

and nutrients to the tissue, being more correlated to the salvageable tissue. Thus, CBF allows us to study which regions are underperfused. Still, both CBV and CBF have some disadvantages: They are heterogeneous between gray and white matter, even in normal tissue; they are susceptible to errors caused by signal clipping, and they are affected if the blood-brain barrier is not intact (Copen et al., 2011; Straka et al., 2010). Moreover, the penumbra in those sequences is underestimated if the bolus is delayed and in short acquisitions. Nevertheless, CBF suffers less from this problem than CBV (Copen et al., 2011; Straka et al., 2010). While the machine learning system selected just a few hidden nodes that compute CBF-related features, it did not select any for CBV, Fig. 7. This may be related with CBF being more related to the penumbra, while CBV appears to identify regions closer to the core. Additionally, the disadvantages of these sequences, pointed out above, may contribute for a more heterogeneous data, hence being harder to capture relations in it. On the contrary, TTP and Tmax are the sequences with the highest importance, according to the squared L2-norm plots of Fig. 7. In fact, these sequences are not directly measuring perfusion, but correlate well with hypoperfusion. TTP and Tmax are more independent of the tissue type (less heterogeneous in gray and white matter) and the acquisition time, than CBV and CBF. Moreover, the lesions are conspicuous in these sequences (Copen et al., 2011; Straka et al., 2010). For this reason, Straka et al. (2010) proposed a method for penumbra estimation based on thresholding the Tmax $>$ 6 s, and further removal of small clusters. Furthermore, the manual segmentation protocol of the SPES database starts by thresholding the Tmax sequence to have a first segmentation of the hypoperfused region. The other MRI sequences are then used to refine it, by removing the sulci, non-stroke pathologies, and previous infarcts (Maier et al., 2017). These considerations show the importance of Tmax for estimating penumbra. The feature maps shown in Fig. 7 appear to identify penumbra patterns, by appearing hypointense in the TTP image and hyperintense on the Tmax, in the area of interest. Some features based on T1c and T2 also appear as important, which may be related to the suppression of sulci, similarly to the manual segmentation protocol. In the spatially distributed explanation of the predictions (Fig. 8), the importance of TTP and Tmax is confirmed. Interestingly, one can note that those two sequences change their importance according to the location. The importance of CBV is not zero everywhere because, although no features are strongly linked to that sequence, there are some residual weights that accumulate during the local interpretability algorithm, even though we employed L1-norm to turn the least important weights to 0. This may be an artifact of the algorithm, although the magnitude of importance of CBV is much lower than the strongest responses, thus can be considered as negligible.

Although the system learned to rely on Tmax and TTP at the expense of CBV and CBF, we know from the literature that the latter sequences should have some discriminative power (Copen et al., 2011; Straka et al., 2010). Still, as previously mentioned, their importance is reduced, as computed by our methodologies. However, upon confirmation with a clinical expert, this may point to a bias introduced during the manual segmentation step, since it heavily relies on the Tmax perfusion sequence that is thresholded at 6 s. The machine learning system may learn to recognize its importance, since the expert similarly applies an intensity threshold on the Tmax image. Moreover, this may account for the success of the top-2 methods in Table 4, since both thresholded Tmax. So, interpreting a model may unveil potentially imperceptible biases on the training data. In fact, the possibility to disclosure problems in the data is pointed out by Ribeiro et al. (2016b) as one of the advantages of developing interpretation methodologies for machine learning systems.

When we trained a system without the CBV and CBF sequences, the results were similar to using all the MRI sequences. Of course, as mentioned before, this may be due to a bias in the manual segmentation procedure. Nevertheless, the interpretation of the system allowed us to identify MRI sequences that are less important, and remove them from the system. Thus, interpretation may help to identify unimportant MRI sequences for some tasks, which can be helpful for reducing acquisition time and cost.

7. Conclusion

In conclusion, we propose a machine learning system based on a RBM as representation mapping and a RF as task-specific learner. Furthermore, we propose methodologies and definitions for the machine learning system interpretability, both globally and locally. Despite being a shallow model, a RBM can still learn meaningful features in an unsupervised way that are useful for segmentation. Indeed, although being unsupervised, it learned to compute tissue specific features, as observed in the feature maps. The fact that it is shallow, however, makes it simpler to find useful information in its weights. This suggests that despite being regarded as “black boxes”, we can still interpret the behavior of these models. We observed that the most important features could extract sequence- and task-specific knowledge. Contrasting with the common belief that these models mix all the information in their weights, in fact these findings suggest that it is employed in an organized fashion. Furthermore, we could verify that the system was able to capture information coherent with expert knowledge, such as the manual segmentation protocol used for BRATS 2013 (Menze et al., 2015) and penumbra estimation for SPES (Maier et al., 2017). This was observed both globally and locally (spatially distributed in the image space). For instance, in the complete tumor vs. normal tissue in BRATS, it was observed through the local interpretability methodology that FLAIR is the most important sequence, as expected. However, for the regions overlapping with the enhancing tumor the system still recognized the importance of the T1c sequence. Also, we could suggest a possible bias towards the importance of the Tmax sequence introduced by the manual segmentation protocol in SPES. With our interpretability methodologies, we aimed at improving the transparency and interpretability of Representation Learning-based methods to increase their acceptance in clinical applications. Finally, we proposed a strategy for feature selection combining RBM features and RF MDI. Summarizing, we present a methodology joining RF and RBM for data understanding, and feature extraction and selection. This approach opens opportunities to understand how MRI sequences are being used for each segmentation task; thus, it can potentially be useful to refine imaging protocols for a given segmentation task, with an impact both in acquisition time and cost. As well, it may be helpful to understand and take advantage of sequence specific features. In the future, we want to investigate how these findings can be extended for other models (for instance, Deep Belief Networks) and applications. Additionally, we will investigate more principled approaches for feature selection using both RBM-MI and RF-MDI.

Acknowledgments

Sérgio Pereira was supported by a scholarship from the Fundação para a Ciência e Tecnologia (FCT), Portugal (scholarship number PD/BD/105803/2014). This work is supported by FCT with the reference project UID/EEA/04436/2013, by FEDER funds through the COMPETE 2020 Programa Operacional Competitividade e Internacionalização (POCI) with the reference project POCI-01-0145-FEDER-006941. This project has received funding from the European Unions Seventh Framework Programme for research,

technological development and demonstration under grant agreement N^{0}600841 and the Swiss National Science Foundation project 205321_169607.

References

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., MÃžller, K.-R., 2010. How to explain individual classification decisions. J. Mach. Learn. Res. 11, 1803–1831.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE T. Neural Netw. 5 (4), 537–550. doi:10.1109/72.298224.

Bennasar, M., Hicks, Y., Setchi, R., 2015. Feature selection using joint mutual information maximisation. Expert Syst. Appl. 42 (22), 8520–8532.

Breiman, L., 2001. Random forests. Mach. learn. 45 (1), 5–32.

Copen, W.A., Schaefer, P.W., Wu, O., 2011. Mr perfusion imaging in acute ischemic stroke. Neuroimaging Clin. N. Am. 21 (2), 259–283.

Cortez, P., Embrechts, M.J., 2011. Opening black box data mining models using sensitivity analysis. In: Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, pp. 341–348.

Craven, M.W., Shavlik, J.W., 1996. Extracting tree-structured representations of trained networks. Adv. Neural Inf. Process. Syst. (NIPS) 24–30.

Criminisi, A., Shotton, J., 2013. Decision Forests for Computer Vision and Medical Image Analysis. Springer Science & Business Media.

Freitas, A.A., 2014. Comprehensible classification models: a position paper. SIGKDD Explor. Newsl. 15 (1).

Gallego-Ortiz, C., Martel, A.L., 2016. Interpreting extracted rules from ensemble of trees: application to computer-aided diagnosis of breast mri. ICML Workshop on Human Interpretability in Machine Learning (WHI) ArXiv:1606.08288.

Ganz, M., Greve, D.N., Fischl, B., Konukoglu, E., 2015. Relevant feature set estimation with a knock-out strategy and random forests. NeuroImage 122.

Hara, S., Hayashi, K., 2016. Making tree ensembles interpretable. ICML Workshop on Human Interpretability in Machine Learning (WHI) ArXiv:1606.05390.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. Hemis: Hetero-modal image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 469–477.

Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Comput. 14 (8).

Hinton, G.E., 2012. Neural Networks: Tricks of the Trade: Second Edition. Springer Berlin Heidelberg.

Hinton, G.E., McClelland, J.L., Rumelhart, D.E., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. MIT Press, Cambridge, MA, USA, pp. 77–109.

Hinton, G.E., et al., 2006. A fast learning algorithm for deep belief nets. Neural. Comput. 18 (7).

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2016. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Med. Image Anal..

Konukoglu, E., Ganz, M., 2014. Approximate false positive rate control in selection frequency for random forest. arXiv:1410.2838.

Krause, J., Perer, A., Ng, K., 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, pp. 5686–5697.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. (NIPS) 1097–1105.

Larochelle, H., Bengio, Y., 2008. Classification using discriminative restricted boltzmann machines. In: Proceedings of the 25th International Conference on Machine learning. ACM, pp. 536–543.

LeCun, Y., et al., 2015. Deep learning. Nature 521 (7553).

Lipton, Z.C., 2016. The mythos of model interpretability. ICML Workshop on Human Interpretability in Machine Learning (WHI) ArXiv:1606.03490.

Louppe, G., et al., 2013. Understanding variable importances in forests of randomized trees. Adv. Neural Inf. Process. Syst. (NIPS).

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Med. Image Anal. 35, 250–269.

McKinley, R., Häni, L., Gralla, J., El-Koussy, M., Bauer, S., Arnold, M., Fischer, U., Jung, S., Mattmann, K., Reyes, M., et al., 2016. Fully automated stroke tissue estimation using random forest classifiers (faster). J. Cerebral Blood Flow Metab..

Meier, R., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2014. Patient-specific semi-supervised learning for postoperative brain tumor segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer International Publishing, pp. 714–721.

Meier, R., Knecht, U., Loosli, T., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2016. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. Sci. Rep. 6.

Meier, R., Porz, N., Knecht, U., Loosli, T., Schucht, P., Beck, J., Slotboom, J., Wiest, R., Reyes, M., 2017. Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma.. J. Neurosurg. 1–9. doi:10.3171/2016.9.JNS16146.

Meier, R., et al., 2014. Appearance-and context-sensitive features for brain tumor segmentation. MICCAI BraTS.

Menze, B.H., Leemput, K.V., Lashkari, D., Riklin-Raviv, T., Geremia, E., Alberts, E., Gruber, P., Wegener, S., Weber, M.A., Szkely, G., Ayache, N., Golland, P., 2016. A generative probabilistic model and discriminative extensions for brain lesion segmentation with application to tumor and stroke. IEEE T. Med. Imaging 35 (4), 933–946.

Menze, B.H., et al., 2015. The multimodal brain tumor image segmentation benchmark (brats). IEEE T. Med. Imaging 34 (10).

Mullins, M.E., Barest, G.D., Schaefer, P.W., Hochberg, F.H., Gonzalez, R.G., Lev, M.H., 2005. Radiation necrosis versus glioma recurrence: conventional MR imaging clues to diagnosis.. Am. J. Neuroradiol. 26 (8), 1967–1972.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning (ICML).

Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436.

Nicolai Meinshausen, P.B., 2010. Stability selection. J. R. Stat. Soc. B 72 (4), 417–473.

Nowozin, S., 2012. Improved information gain estimates for decision tree induction. In: International Conference on Machine Learning (ICML).

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of mri scale standardization. IEEE T. Med. Imaging 19 (2), 143–150.

Olden, J.D., Jackson, D.A., 2002. Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. 154 (1).

Pedregosa, F., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE T. Pattern Anal. 27 (8), 1226–1238.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in mri images. IEEE T. Med. Imaging 35 (5), 1240–1251.

Pereira, S., Pinto, A., Oliveira, J., Mendrik, A.M., Correia, J.H., Silva, C.A., 2016. Automatic brain tissue segmentation in mr images using random forests and conditional random fields. J. Neurosci. Meth. 270, 111–123.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Model-agnostic interpretability of machine learning. ICML Workshop on Human Interpretability in Machine Learning (WHI) ArXiv:1606.05386.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016, pp. 1135–1144.

Salakhutdinov, R., Mnih, A., Hinton, G., 2007. Restricted boltzmann machines for collaborative filtering. In: International Conference on Machine Learning (ICML). ACM, pp. 791–798.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034.

Smolensky, P., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. MIT Press, Cambridge, MA, USA, pp. 194–281.

Straka, M., Albers, G.W., Bammer, R., 2010. Real-time diffusion-perfusion mismatch analysis in acute stroke. J. Magn. Reson. Imaging 32 (5), 1024–1037.

Szegedy, C., Inc, G., Zaremba, W., Sutskever, I., Inc, G., Bruna, J., Erhan, D., Inc, G., Goodfellow, I., Fergus, R., 2014. Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR).

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B. Met. 267–288.

van Tulder, G., de Bruijne, M., 2015. Why does synthesized data improve multi-sequence classification? In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer.

van Tulder, G., de Bruijne, M., 2016. Combining generative and discriminative representation learning for lung ct analysis with convolutional restricted Boltzmann machines. IEEE T. Med. Imaging 35 (5), 1262–1272.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: Improved n3 bias correction. IEEE T. Med. Imaging 29 (6), 1310–1320.

Tzeng, F.Y., Ma, K.L., 2005. Opening the black box - data driven visualization of neural networks. In: VIS 05. IEEE Visualization, 2005., pp. 383–390.

Vergara, J.R., Estévez, P.A., 2014. A review of feature selection methods based on mutual information. Neural Comput. Appl. 24 (1), 175–186.

Wang, S., Summers, R.M., 2012. Machine learning and radiology. Med. Image Anal. 16 (5), 933–951.

Wejchert, J., Tesauro, G., 1989. Neural network visualization.. Adv. Neural Inf. Process. Syst. (NIPS) 465–472.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV). Springer, pp. 818–833.

Zhen, X., Wang, Z., Islam, A., Bhaduri, M., Chan, I., Li, S., 2016. Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Med. Image Anal. 30, 120–129.

Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: a quantitative study. Artif. Intell. Rev. 22 (3), 177–210.

Zrihem, N.B., Zahavy, T., Mannor, S., 2016. Visualizing dynamics: from t-sne to semi-mdps. ICML Workshop on Human Interpretability in Machine Learning (WHI) ArXiv:1606.07112.