On Feature Relevance in Image-based Prediction Models: An Empirical Study

Ender Konukoglu[1⋆], Melanie Ganz[1⋆], Koen van Leemput[1,2,3], and Mert R. Sabuncu[1] for the Alzheimers Disease Neuroimaging Initiative ⋆⋆

[1] Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, USA
[2] Department of Applied Mathematics and Computer Science, DTU, Denmark
[3] Departments of Information and Computer Science and of Biomedical Engineering and Computational Science, Aalto University, Finland

Abstract. Determining disease-related variations of the anatomy and function is an important step in better understanding diseases and developing early diagnostic systems. Machine-learning based medical image analysis methods provide valuable tools to determine such variations. In particular, image-based multivariate prediction models and "relevant features" they produce are attracting attention from the community. In this article, we present an empirical study on the relevant features produced by two recently developed discriminative learning algorithms: neighborhood approximation forests (NAF) and the relevance voxel machine (RVoxM). Specifically, we examine whether the sets of features these methods produce are exhaustive; that is whether the features that are **not** marked as relevant carry disease-related information. We perform experiments on three different problems: image-based regression on a synthetic dataset for which the set of relevant features is known, regression of subject age as well as binary classification of Alzheimer's Disease (AD) from brain Magnetic Resonance Imaging (MRI) data. Our experiments demonstrate that aging-related and AD-related variations are widespread and the initial sets of relevant features discovered by the methods are not exhaustive. Our findings show that by knocking-out features and re-training models, a much larger set of disease-related features can be identified.

1 Introduction

Image-based prediction models hold great promise for improving clinical practice. The ability to predict the state of a disease based on its anatomical and

⋆ E. Konukoglu and M. Ganz contributed equally to this work.

⋆⋆ Data used in preparation of this article was obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (http://www.adni-info.org/). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

functional signatures opens up new avenues for early diagnostic systems. To this end, great progress has been made in developing accurate prediction tools. In particular, multivariate discriminative learning models have been successfully employed both for image-based classification and regression.

Furthermore, medical images provide in-vivo observations that reveal the disease-related variations of the anatomy and function. Comparing images of healthy controls and patients can highlight such variations on a macro-scale that would be difficult to identify using histopathology. This is essential for improving our understanding of disease and refining the predictive power of learning methods.

Supervised image-based prediction models have started to play an important role in determining regions that exhibit disease-related variations. These models determine a set of regions, also referred to as "relevant features", on which measurements are pertinent to achieve good prediction accuracy. The statistical significance of the predictions, i.e. that they are different than a random guess, suggests that such sets of relevant features indeed carry information about disease-specific changes. Although this is a good starting point, an important question remains: Do the methods produce exhaustive sets? That is, does the set of relevant features reflect most of the variation associated with the disease-specific effects? Recent work has focused on constructing prediction models aiming to determine relevant sets [1] and developing efficient statistical tests for creating maps of relevant regions [2]. To the best of our knowledge, no previous work has analyzed the question whether the set of regions **not** chosen as "relevant" by prediction models exhibit any disease-specific changes. Here, we empirically study this issue.

This article presents a set of experiments with two different image-based prediction models: neighbourhood approximation forests (NAF) [3] and the relevance voxel machine (RVoxM) [1]. Both of these methods have been used to perform image-based classification and regression; both methods offer a way to quantify feature relevance for prediction. We designed an experimental setup that aims to answer whether the relevant features provided by these models form an exhaustive set. Using this setup we performed experiments on three different datasets: synthetic, OASIS [4] and ADNI[4]. The first two are image-based regression tasks, while the last one is a classification task. We present the experimental setup, briefly describe the methods and present experimental findings.

2 Experimental Design

Our experimental design relies on the following observation: if the set of relevant features is exhaustive, then removing it from the pool of measurements should result in a set of features that contains no information about the disease. Consequently, an image-based prediction model trained on this non-informative set, should be as predictive as a random prediction model, i.e. prediction should not

[4] For detailed information, visit http://www.adni-info.org/

```
Feature Set = Entire feature set of measurements;
iteration = 1;
while iteration < Total Number of Iterations && #(Feature Set) != 0 do
    Train on the training set using Feature Set;
    Compute accuracy on the test set;
    Perform random permutation tests on the trained prediction model;
    Compute the most relevant features;
    Knockout - remove the most relevant features from Feature Set;
    iteration++;
end
```

Algorithm 1: Pseudocode for the experimental design

be statistically significant. Based on this consideration, we designed an iterative setup. At the first iteration, we learn a prediction model using a training dataset and the entire set of measurements available. We first record the relevant features suggested by the method and then compute the model's accuracy on a separate test dataset. We then perform random permutation tests as suggested in [5]; we randomly permute the labels in the test set and recompute the prediction accuracies using the trained model. Based on the random permutation tests, we compute the statistical significance of the prediction model at that iteration, i.e. percentage of random permutations achieving better accuracy than the trained model. For the second iteration we "knock-out" the relevant features, i.e. remove them from the feature set, and **retrain** the model on the remaining set of measurements. Then we repeat the same procedure multiple times. Algorithm 1 presents the pseudocode for this procedure.

3 Materials

3.1 Algorithms

We represent images as feature vectors $\mathbf{x}$, where each $\mathbf{x}$ is a concatenation of measurements taken from different locations within the image, e.g. intensities. The quantities to predict are denoted as y for the regression tasks and as c for the binary classification tasks. Both methods we consider are supervised; they use a training database $\{\mathbf{x}_n, (y, c)_n\}_1^N$ to learn the prediction model. Below, we briefly describe each method.

Neighbourhood Approximation Forests (NAF) Neighbourhood approximation forests [3], a variant of random decision forests [6], is a supervised discriminative learning method that uses tree-based approximate nearest neighbour search. During the training phase NAF learns sequences of axis aligned binary tests on the feature space that approximate the neighbourhood structure induced by a user-defined distance on the labels. Once learning is complete, for a test image $\mathbf{x}$, NAF computes its predicted label by identifying the "closest" M training images and using their labels. The prediction model combines these

labels by weighting them based on the approximate affinities of the training images to the test one as computed by NAF. As a decision forest based system, various feature relevancy measures can be used with NAF [7, 8]. Here, we adopt "selection frequency", freq($\mathbf{x}_i$), which basically counts the number of times each element $\mathbf{x}_i$ is used within the learned forest. Ordering freq($\mathbf{x}_i$) provides a relative "importance" measure to each element in $\mathbf{x}$. For further details on NAF, its parameters and the selection frequency we refer the reader to [3] and [7, 8].

Relevance Voxel Machine (RVoxM) Relevance voxel machine [1], a variant of relevance vector machine (RVM) [9], is a sparse learning algorithm. The prediction model is linear: $y(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{M} w_i x_i = \mathbf{w}^T \mathbf{x}$, were $\mathbf{w}$ is a weight vector that is optimally sparse. In RVoxM, the algorithm learns $\mathbf{w}$ for a training dataset by optimizing a cost function that is a combination of the data likelihood and a prior model that encourages sparsity and spatial clustering of features. RVoxM is applied to regression tasks as $\hat{y} = \mathbf{w}^T \mathbf{x}$ and to binary classification tasks as $p(c = 1) = \sigma(\mathbf{w}^T \mathbf{x})$, where $\sigma(\cdot)$ is a sigmoid function mapping scalars to class probabilities. In the training phase, RVoxM computes the optimal sparse $\mathbf{w}$ vector. During prediction, only the corresponding elements in the feature vector $\mathbf{x}$ contribute and therefore, these elements are the "relevant" features of the model. For further details on RVoxM we refer the reader to [1].

3.2 Data

Synthetic Data - Image-based regression In the first experiment, we used NAF and RVoxM on a synthetic dataset for an image-based regression problem, where the ground truth set of relevant pixels is known. Our aim with this dataset is to test the validity of the experimental design described in Section 2. The dataset contains images synthetically generated from an underlying natural image. Each image is composed of 121 7×7 non-overlapping patches, where the intensities of each patch are computed by multiplying a 2D Gaussian kernel centered at the patch center with the intensities of the underlying image. The magnitude of each kernel is drawn uniformly from the interval $[-1, 1]$. We further add normally distributed iid white noise to the images to achieve an SNR of 20dB. For each image we assign the kernel magnitude of the 50^{th} patch as the variable of interest. Figure 1 shows the relevant pixels, some example images and the corresponding scalar values. The feature vector $\mathbf{x}$ for each image contains the intensity values of all pixels, i.e. 5929 values. In this fashion, we constructed 10 different datasets each containing 400 training and 40 test images.

OASIS - Image-based regression In the second experiment, we used the methods for image-based regression of subject age on the publicly available cross-sectional Open Access Series of Imaging Studies (OASIS) [4]. We used structural brain MRI (T1-weighted) scans of 414 healthy subjects that were processed with the FreeSurfer software suite [10, 11] and transferred to a common coordinate system via affine registration. From the skull-stripped and normalized brains at resolution $2 \times 2 \times 2$ mm^3, we randomly sampled 15,613 intensity values for computational efficiency and used those as our features. We randomly chose 100

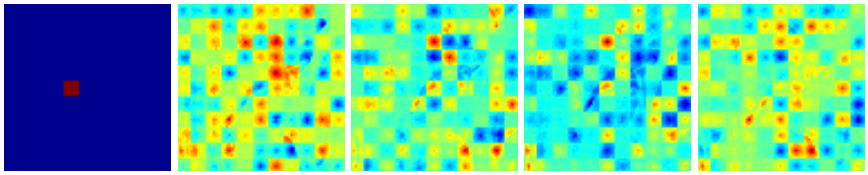

Fig. 1. Images from the synthetic dataset. Left-most: ground truth relevant pixels. Right: four example images with their corresponding scalar variables.

of the 414 subjects as a test set and used the remaining 314 for training; we repeated this procedure ten times.

ADNI - Image-based classification In the last experiment we used NAF and RVoxM for image-based classification of Alzheimer's Disease (AD)of the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset[5]. Our analysis used MRI scans from 180 AD patients and an age and sex-matched group of 180 controls. We processed all MRI scans with the FreeSurfer software suite and computed subject-specific models of the cortical surface as well as thickness measurements across the entire cortical mantle [12]. Subject-level thickness measurements were then transferred to a common coordinate system, via a surface-based nonlinear registration procedure [13]. For computational efficiency, we utilized the left-hemisphere of the *fsaverage5* representation, consisting of 10,242 vertices. We smoothed the cortical thickness maps with a Gaussian kernel with a full-width-half-maximum of 5 mm. To evaluate our algorithms, we conducted a ten-fold cross-validation. First we divided the data into 10 subgroups (of 18 patients and 18 controls); then during each fold, one patient and one control subgroup were set aside as the test set, while the rest of the data were used for training.

4 Results and Discussions

On the datasets described above we performed the iterative experiment as explained in Section 2. For RVoxM we let the algorithm converge during the training phase, which produced a set of relevant features. At each iteration of the experiment these relevant features were knocked-out. For NAF, the selection frequency provided an ordering for all features. At each iteration of the experiment, we knocked out the top 3% (1% for the synthetic dataset) of the features with respect to the selection frequency ordering. For each experiment, at each iteration we computed: i) prediction accuracies, ii) 1000 random permutation tests (100 permutations for each fold) and iii) p-values of the prediction accuracies with respect to the permutation tests. As prediction accuracies, we used root-mean-square error (RMSE) for the regression tasks and misclassification

[5] For detailed information, visit http://www.adni-info.org/

rate for the classification task. The p-values were computed with respect to the average accuracy obtained over different folds of the same experiment.

Figure 2 summarizes the results. For each experiment we provide two graphs. The first one plots the prediction accuracy with respect to % of features that were knocked-out. The solid lines are the average accuracies obtained in our 10 fold experiments and the error bars correspond to standard deviations and standard errors. The same plot also presents the average statistics on the accuracies obtained during the random permutation tests. The second graph plots the % of features knocked-out vs. the p-values with respect to random permutation tests.

The results for the experiments on the synthetic dataset (first column in Figure 2) demonstrate the validity of our experimental design. After the first few iterations, the trained prediction models for both algorithms behave the same way as random permutation tests. This means both algorithms identified all the relevant pixels in the first few iterations and in the latter iterations the models are trained on non-informative pixels only. To validate this, we computed the least common set of relevant pixels at each iteration by intersecting the sets of knocked-out pixels over the different folds. Figure 3(a) plots the % of ground truth (GT) relevant pixels identified by the least common feature sets. Indeed we see that both algorithms discover all the relevant pixels and after that point both prediction models lose their significance.

The experimental results for the age regression problem on the OASIS dataset display a different behavior than the synthetic experiments. The prediction model remains significant ($p < 0.01$) until the point, where we knocked-out 96% of the total number of features for NAF and 83% of the features for RVoxM. This indicates that over 80% of the features might carry information for predicting subject age, which is not surprising since the effect of age is expected to be widespread. Furthermore, the accuracies of the prediction models decrease with each knock-out suggesting that the algorithms can make better use of some regions in the brain to perform predictions than others.

The results for the AD classification problem on the ADNI dataset show a yet different type of behavior. The prediction models remain significant until 66% of the features are knocked-out for NAF and 75% for the RVoxM. This suggests that the effect of AD is also widespread across the cortex. But in contrast to the age regression problem, the accuracy changes faster in the first iterations of the experiment. This suggests that there is a small set of features, which the algorithms can exploit substantially better for prediction. The difference is striking between the first and the second iteration of the experiments using NAF.

The plots given in Figure 2 demonstrate that the set of relevant features suggested by image-based prediction algorithms, such as RVoxM, might not be exhaustive. For NAF and other forest-based methods, there is no obvious way to determine a set of relevant features without using an ad-hoc threshold like the 3% value we used. Nevertheless, observing the plots we can say that the set of most frequently used features in NAF might not capture the entire set of relevant ones. In Figures 3(b) and (c) we visualize this behavior for NAF and the

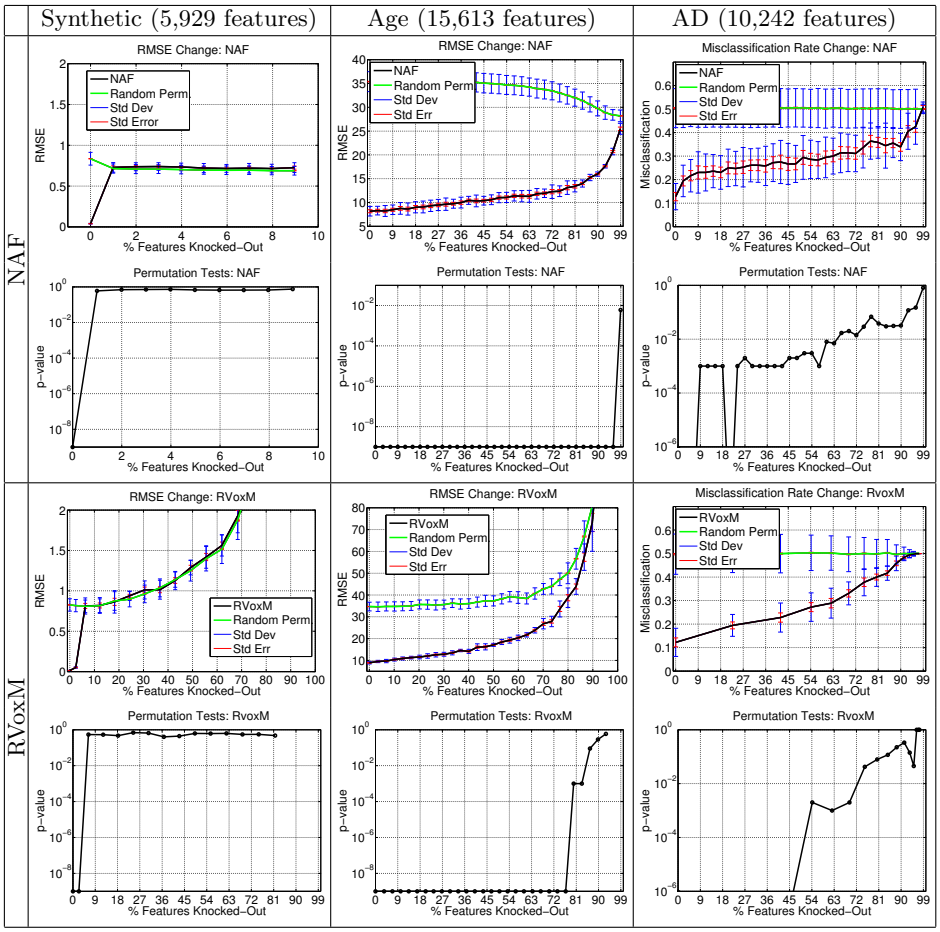

Fig. 2. Graphs show how prediction models change with respect to % of feature knock-out. For each experiment we plot two graphs: i) accuracy change and ii) p-values with respect to random permutation tests. Results for synthetic experiments demonstrate that both models identify the grunt truth relevant features in the first few iterations and after that point prediction model becomes similar to random prediction ($p > 0.5$). Results for age regression and AD classification demonstrate that majority of the features are informative about the effect of interest. Only after knocking out more than 70% of the features p-values for the prediction models become $p > 0.05$.

RVoxM, respectively. The top rows display the regions which were suggested as relevant by the algorithms at the first iteration of the experiment. The bottom rows display the regions which were knocked-out to achieve prediction models that are not significant anymore. We would like to emphasize that these latter images are not necessarily displaying exhaustive sets either.

One important aspect our experiments have shown is the importance of knocking-out features and retraining. Image-based prediction models aim to

achieve good accuracies; and not necessarily to capture the entire set of relevant features. As a result, it is not trivial to interpret the features suggested as relevant by these methods. Thus, we believe the knock-out/retrain strategy is crucial in understanding feature relevance for image-based prediction models.

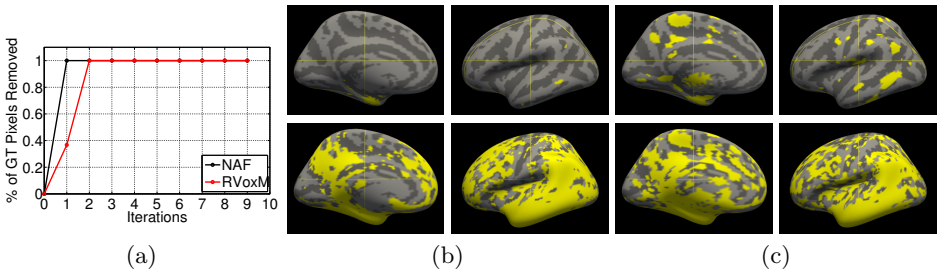

Fig. 3. (a) % of ground truth relevant pixels identified by the methods in the synthetic experiments. (b) Knocked-out features in NAF - in the first iteration (top) and until the point NAF prediction became non-significant (bottom). (c) Same maps for RVoxM.

5 Conclusions

In this article, we presented an empirical study of feature relevance for image-based prediction models. The main question we wanted to answer was whether sets of relevant features suggested by image-based prediction models form an exhaustive set of all features that carry information on disease-specific changes. Our experimental findings demonstrate that this is not necessarily the case in aging and AD. Furthermore, our results show that the knock-out/retrain strategy employed in our experimental design can be critical in understanding and interpreting feature relevance for image-based prediction models.

6 Acknowledgments

This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. This work also involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program and/or High-End Instrumentation Grant Program; specifically, grant number(s) S10RR023401, S10RR019307, S10RR019254 and S10RR023043. The research was also supported by the Alfred Benzon and the Lundbeck Foundation.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

References

1. Sabuncu, M., Van Leemput, K.: The relevance voxel machine (rvoxm): A self-tuning bayesian model for informative image-based prediction. Medical Imaging, IEEE Transactions on **31**(12) (2012) 2290–2306
2. Gaonkar, B., Davatzikos, C.: Deriving statistical significance maps for svm based image classification and group comparisons. In: MICCAI 2012. Springer (2012) 723–730
3. Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A.: Neighbourhood approximation using randomized forests. Medical Image Analysis (2013)
4. Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience **19**(9) (2007) 1498–1507
5. Good, P.I.: Permutation, parametric and bootstrap tests of hypotheses. Springer Science+ Business Media (2005)
6. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural computation **9**(7) (1997) 1545–1588
7. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32
8. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics **8**(1) (2007) 25
9. Tipping, M.: The relevance vector machine. In: Advances in Neural Information Processing Systems 12, MIT Press (2000) 652 – 658
10. Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage **9**(2) (1999) 179–194
11. Fischl, B., Sereno, M.I., Dale, A.M.: Cortical surface-based analysis ii: Inflation, flattening, and a surface-based coordinate system. Neuroimage **9**(2) (1999) 195–207

12. Fischl, B., Dale, A.M.: Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences **97**(20) (2000) 11050–11055
13. Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., et al.: High-resolution inter-subject averaging and a coordinate system for the cortical surface. Human brain mapping **8**(4) (1999) 272–284