Title of the Paper

Michael Mueller, Jan Riedo, Michael Rebsamen

***Abstract*—Machine learning (ML), white matter (WM), grey matter (GM), cerebrospinal fluid (CF), background (BG)**

***Index Terms*—MRI, Segmentation, Machine Learning, DF, kNN, SVM**

I. Introduction

Segmentation of brain tissues from magnetic resonance images (MRI) has many clinical applications. Clinicians gain useful information from a separation of tissue into its three main anatomical types: white matter, grey matter, and ventricles. However, manual segmentation of MRI is a labour-intensive task requiring expert skills. Fully automatic approaches for brain tissue segmentation are therefore a topic of active research. A good algorithm classifies the tissue types with high accuracy across a variety of images from different patients. Such a classification is a typical task for machine learning. These algorithms tend to perform well given enough training data during the learning phase. The availability of ground-truth data in sufficient quantity and quality for supervised learning is a particular challenge when working with medical images due to privacy concerns and the costs for manual segmentation. Optimization of the learning phase with a limited number of training data is therefore required.

FIXME: kNN is a popular classification method for MR data and has successfully been applied in MR brain segmentation [1]–[3]

FIXME: Base paper on df [4].

II. Methods

A. Dataset

All experiments were conducted on a subset of 100 unrelated subjects from a dataset provided by the *Human Connectome Project* [5]. From each individual, a total of eight 3-tesla head MRI are available: T1 and T2-weighted image volumes not skull-stripped (but defaced for anonymization) and skull-stripped with a bias field correction, and both modalities once in native T1 space and once in MNI-atlas space [6].

Ground-truth labels are automatically generated using *FreeSurf*, assigning each voxel either to background, white matter, grey matter, or ventricles. The dataset was split in a training set with 70 images and a test set with 30 images.

B. Pipeline

Training and testing data are loaded sequentially, each put through the pipeline consisting of: registration, pre-processing, feature extraction and ML training/classification. For testing, two additional steps, namely post-processing and evaluation are added.

Firstly, the data is loaded and registered to an atlas with a multi-modal rigid transformation using a regular step gradient descent optimizer. FIXME: In the preprocessing module skull stripping and bias field correction are applied in order to have images of the brain only, with less influence of the MRI scanning characteristics. Furthermore, a gradient anisotropic diffusion filter and z-score normalization is applied.

Preprocessed data is then fed into the feature extraction module, where seven features are computed. The feature matrix consists of three coordinate features, a T1 and a T2 intensity feature, and a T1 and T2 gradient feature. During feature extraction, a random mask is applied in order to randomly select a fraction of the voxels available. The mask is adjustable individually for BG, WM, GM, and CF. This is where the pathways of training and testing split up: training data is lastly fed to a certain supervised machine learning algorithm for training, whereas the testing data is classified with the previously created model. The classified testing data is then forwarded to a post-processing module where a dense conditional random field [7] is applied. Finally, the classification is evaluated based on a comparison with the ground truth, where a dice coefficient is computed (see chap. II-E).

FIXME: The medical image analysis pipeline consists of seven phases. In the pre-processing phase, an intensity normalization is performed on the images to have comparable grey scale ranges among all images in the subsequent steps. Similarly, the registration phase registers the images to an atlas space to have a common coordinate system. In the feature extraction phase, seven features are calculated for each voxel: three coordinates corresponding to the position of the voxel in the atlas space, and an intensity and a gradient on both the T1 and T2 modalities. A subset of voxels is choosen randomly to be used for training. During the learning phase, the selected algorithm is trained to classify voxels based on the given features. In the segmentation phase, the trained algorithm is used to classify all voxels in previously unseen test images. The post-processing phase aims to reduce noise by eliminating small isolated regions with a dense conditional random field [7]. Finally, the evaluation phase assesses the performance of the segmentation by comparing the result to ground-truth and calculating a dice coefficient for each class.

C. Training

TODO: Describe training of machine learning algorithms
TODO: **Short** intro to used algorithms?

D. Support Vector Machine (SVM)

Classification using Support Vector Machines (SVM) tries to find a hyperplane separating the data in a high-dimensional

Biomedical Engineering, University of Bern
Authors e-Mail: michael.mueller4@students.unibe.ch, jan.riedo@students.unibe.ch, michael.rebsamen@students.unibe.ch

feature space. Given the feature vector x_i and the binary label y_i, the SVM solves the following optimization problem during training:

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \\ \text{s.t. } & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \ i = 1, \ldots, m \\ & \xi_i \geq 0, \ i = 1, \ldots, m \end{aligned} \tag{1}$$

where w is the normal vector and b the offset of the separating hyperplane and $\phi(x_i)$ maps x_i into a higher-dimensional space.

The SVM implementation is based on libSVM [8]. Multi-class classification is solved with a *one-against-one* approach. To output probabilities, the predictions are calibrated using *Platt* scaling in which multiclass problems require an additional cross-validation which is an expensive operation for large datasets.

Given the relative low number of available features, we have chosen a radial basis function (RBF) kernel. A regularization term C and a model capacity γ needs to be chosen. These hyperparameters were determined with an exhaustive search and cross-validated on a subset of the training data, yielding $C = 500$ and $\gamma = 0.00005$.

E. Performance Evaluation

The Dice coefficient is a commonly used metric to compare the spatial overlap, ranging from 0 (no overlap) to 1 (perfect overlap). To evaluate the accuracy of the segmentation, a Dice coefficient is calculated between the prediction (E) and ground-truth (G) for each of the three labels.

$$D = \frac{2|E \bigcap G|}{|E| + |G|} = \frac{2TP}{2TP + FP + FN} \tag{2}$$

F. Infrastructure

TODO: Describe UBELIX, libraries

III. Results

All algorithms tested and optimized were able to yield a good result for brain segmentation. The performance measured with the dice coefficient can be found in Tab. (I). Comparisons of computation time can be seen in Tab. (II).

A. Ground Truth Validity

The importance of looking at the images and not only at the numbers shall be presented based on one example MRI image, segmented with kNN. In Fig. (1) we see one slice of a brain, segmented in three different ways. On the left, we see a kNN segmentation based on non-coordinate features. The one in the middle is segmented with kNN based on all features. On the right we see the ground truth. Although the middle image shows a substantially better result in ventricle dice than the left (0.68 vs. 0.76) we can barely see an improved ventricle segmentation with bare eyes. What we see however, is a big difference on how detailed the white and grey matter are segmented, thus leading to a better overall segmentation. Another fact which shall be presented here is that the ground truth image is not a real ground truth. It is an image segmented by another algorithm. In Fig (1a) a better segmentation of the center part (white matter) is achieved compared to the ground truth image in Fig (1c) (background).

[Fig. 1 about here.]

B. Feature Inspection

Feature selection is another key part of the machine learning process. Features are also called variable or attribute and describe the model. Irrelevant and redundant features do not contribute to the accuracy of the predictive model, at worst they decrease the accuracy. The used feature set consists of seven features, f1-f3: Coordinate features, f4: T1 intensity, f5: T1 gradient, f6: T2 intensity, f7: T2 gradient. The following Fig (2) shows the scatter matrix of all the features. On the diagonal are the histograms for each feature. The right upper part of the diagonal visualizes the linear correlation between each of the feature with the associated correlation coefficient. The left bottom part of the diagonal is redundant to the upper part. There is a moderate uphill relationship for the feature f4 & f5, f5 & f6, f5 & f7. A strong uphill linear relationship for coefficients over 0.7, in this case for the feature f6 & f7. This imply, that only the the first three feature, the coordinates, are independent.

[Fig. 2 about here.]

[Fig. 3 about here.]

[Fig. 4 about here.]

C. Random Mask Optimization

One major task to handle was the low value for the ventricles. Dice values above 0.5 were hard to achieve. One way to improve the dice for ventricles was to optimize the random mask with respect to the fraction of ventricle voxels taken into account. The effects of the random mask on the ventricle dice can be seen in Fig. (5). Best results were achieved with a fraction of 0.004 ventricles, approximately the same fraction as for white matter and grey matter. All following results are based on this optimized mask.

[Fig. 5 about here.]

D. Algorithm Performance

The decision forest algorithm was enhanced with normalized features, a higher number of ventricle voxels in the training set and the optimization of the hyperparameters (see Fig. V). With this settings, the max dice coefficient was lifted from 0.703 to 0.754. This result was achieved with 80 trees and 3000 max nodes.

[Fig. 6 about here.]

Statistical distribution of the dice coefficients can be seen in Fig. V. DF and SVM achieve a similar mean dice score but SVM has a lower variance for the ventricles.

[Fig. 7 about here.]

Comparison of computation time for training and testing is shown in Fig. V.

[Fig. 8 about here.]

[TABLE 1 about here.]

[TABLE 2 about here.]

IV. Discussion

TODO: feature importance? which algorithm to choose for which use-case?

The linear increase of training time with the size of training data in SGD is due to the fixed amount of iterations used. One could either lower the number of iterations inversely proportional to the size of training data or use a threshold on the loss function for early termination instead.

We have observed a rather small influence of the size of the training set, DF and SVM reaching a similar dice coefficient with either 3 or 70 training samples.

V. Conclusion

The major challenge in the current setup remains the quality of ground-truth data. As long as the test set is the output from an other (imperfect) algorithm, any approach is just an approximation of the other mechanism.

Acknowledgement

Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

References

[1] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," *NeuroImage*, vol. 21, no. 3, pp. 1037–1044, mar 2004.

[2] C. A. Cocosco, A. P. Zijdenbos, and A. C. Evans, "A fully automatic and robust brain MRI tissue classification method," *Medical Image Analysis*, vol. 7, no. 4, pp. 513–527, dec 2003.

[3] S. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, template moderated, spatially varying statistical classification," *Medical Image Analysis*, vol. 4, no. 1, pp. 43–55, mar 2000.

[4] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

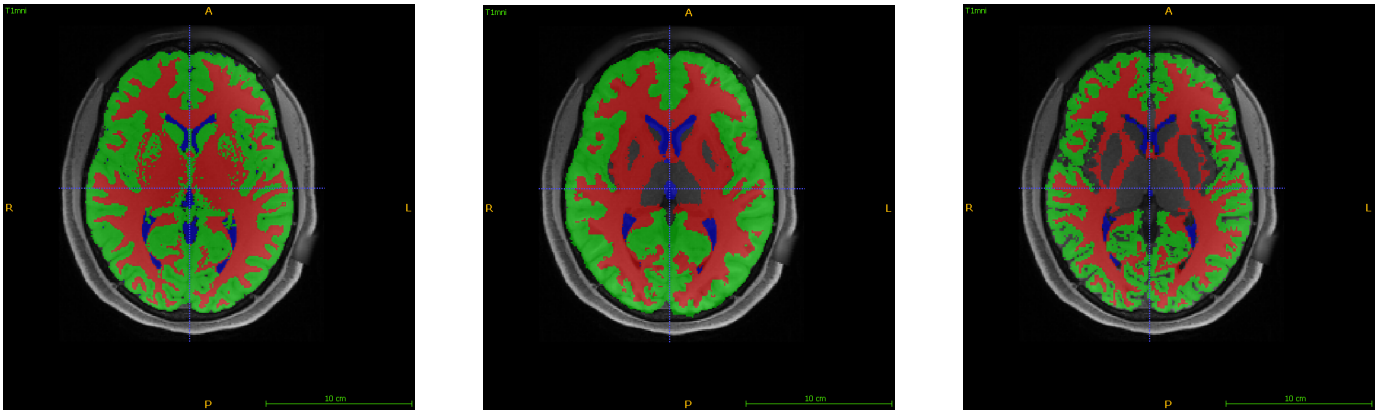
[5] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[6] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike *et al.*, "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm)," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[7] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[8] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

LIST OF FIGURES



(a) (b) (c)

Fig. 1. (a) kNN segmented image based on non-coordinate features with dice values 0.86/0.82/0.68. (b) kNN segmented image based on all features with dice values 0.80/0.79/0.76. (c) Ground truth.

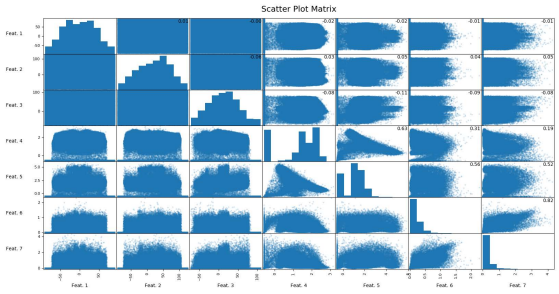


Fig. 2. Scatter plot of the features with correlation coefficient

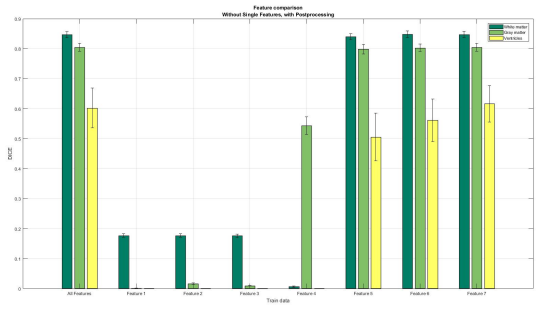


Fig. 3. Feature evaluation with Decision Forest by removing a single feature and with preprocessing

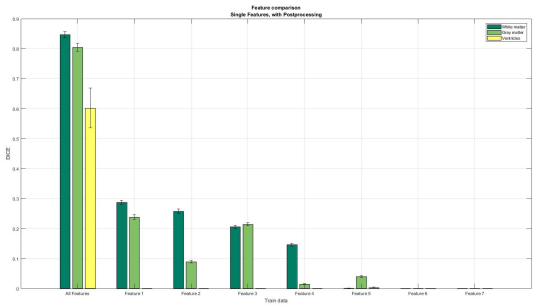
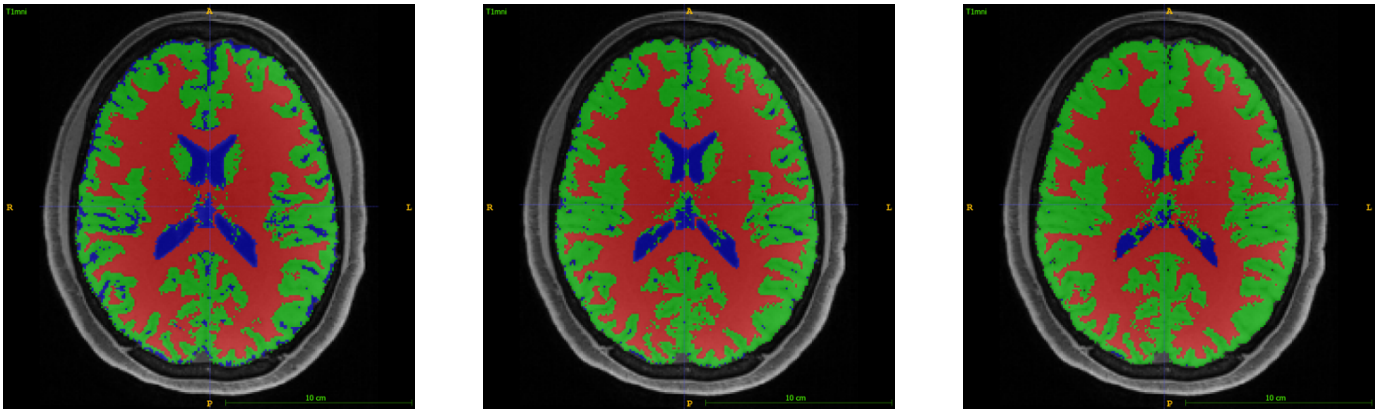


Fig. 4. Feature evaluation with Decision Forest by using a single feature and with preprocessing



(a) (b) (c)

Fig. 5. Optimization of the random mask parameter for ventricles. Fraction of ventricle voxels taken into account f_v and dice value for this certain parameter d_v are: (f_v / d_v) (a) 0.4 / 0.22, (b) 0.04 / 0.44, and (c) 0.004 / 0.62.

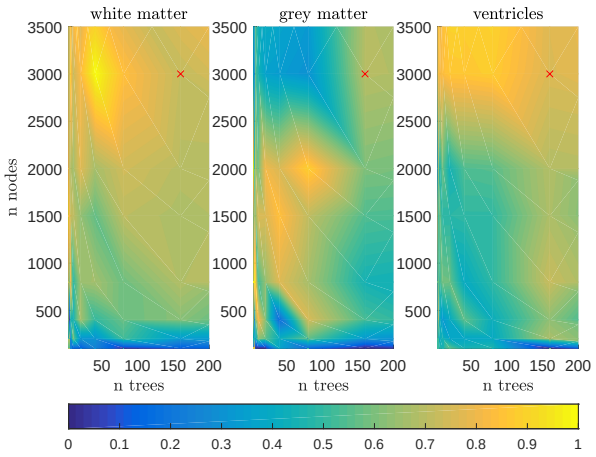


Fig. 6. DF plot of grid search for white matter, grey matter and ventricles. The red cross marks the chosen hyperparameters number of trees = 160 and maximum nodes per tree = 3000. Color does not represent dice, the data is stretched individually for all three plots.

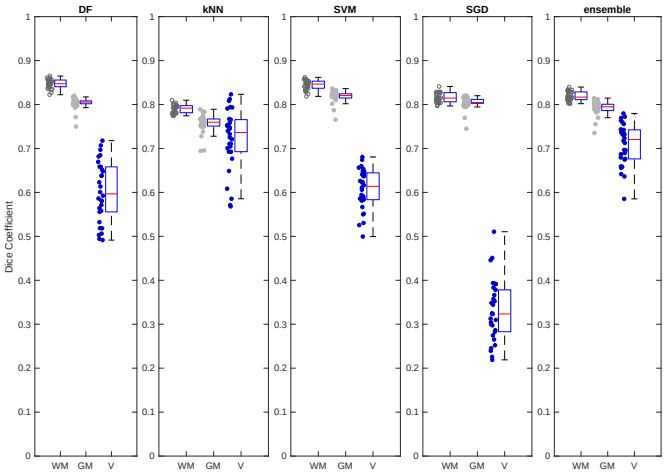


Fig. 7. Distribution of dice coefficients with optimal hyper-parameters for each algorithm on the full training set of 70 images.

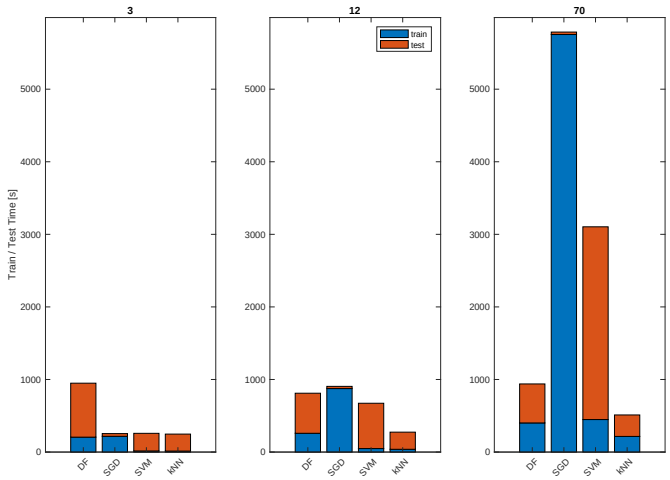


Fig. 8. Time for training and testing of the algorithms with training set sizes of 3, 12 and 70 samples. Test time is for one sample.

List of Tables

TABLE I
PERFORMANCE COMPARISON OF ML ALGORITHMS

Features	Size Dataset	DF	GMM	kNN	SGD	SVM	ensemble
All (f1-f7)	3	0.85/0.81/0.62	-	0.70/0.57/0.50	0.82/0.80/0.35	0.83/0.80/0.61	-
	12	0.85/0.81/0.59	-	0.75/0.66/0.67	0.82/0.80/0.33	0.84/0.81/0.61	-
	70	0.85/0.80/0.60	-	0.80/0.76/0.72	0.82/0.80/0.33	0.84/0.82/0.61	0.82/0.79/0.71
Coordinates only (f1-f3)	3	0.67/0.63/0.22	-	0.70/0.55/0.41	0.17/0.23/0.00	0.59/0.52/0.0	-
	12	0.67/0.64/0.11	-	0.74/0.63/0.56	0.19/0.22/0.00	0.59/0.57/0.0	-
	70	0.67/0.64/0.16	-	0.77/0.71/0.62	0.17/0.21/0.00	0.60/0.58/0.31	-
All non-coordinates (f4-f7)	3	0.84/0.80/0.50	-	0.85/0.80/0.45	0.82/0.80/0.34	0.84/0.79/0.0	-
	12	0.85/0.80/0.49	-	0.85/0.81/0.45	0.82/0.80/0.33	0.85/0.80/0.45	-
	70	0.85/0.80/0.48	-	0.85/0.81/0.54	0.82/0.80/0.34	0.85/0.80/0.44	-

Overview of achieved accuracy for the different algorithms. Mean dice scores for white matter/grey matter/ventricles.
f1-f3: Coordinate features, f4: T1 intensity, f5: T1 gradient, f6: T2 intensity, f7: T2 gradient.

TABLE II
RUNTIME

Features	Size Dataset	DF	GMM	kNN	SGD	SVM
All (f1-f7)	3	205.4/22310.2	-	13.4/7023.7	216.9/1126.5	15.1/7289.7
	12	258.7/16563.6	-	38.1/7090.0	875.0/903.8	48.2/18730.5
	70	401.4/16116.2	-	215.5/8873.5	5753.3/1010.6	448.1/79668.4
Coordinates only (f1-f3)	3	-	-	10.4/4391.5	-	15.4/12178.7
	12	-	-	34.7/5449.3	-	62.2/43404.3
	70	-	-	196.4/6112.8	-	957.9/221440.5
All non-coordinates (f4-f7)	3	-	-	10.1/10084.7	-	12.4/6647.9
	12	-	-	34.6/18768.6	-	39.8/18691.1
	70	-	-	194.2/16555.7	-	323.2/80532.7

FIXME: Overview of the computation time in seconds for all algorithms (training time/testing time). Computation time includes pre- and post-processing.