

Random Subspace Ensemble Artificial Neural Networks for First-episode Schizophrenia Classification

Roman Vyškovský
 Masaryk University, Research
 Centre for Toxic Compounds in
 the Environment (RECETOX),
 Kamenice 3, 625 00 Brno,
 Czech Republic
 Email:
 vyskovsky@recetox.muni.cz

Daniel Schwarz, Eva Janoušová
 Masaryk University, Institute of
 Biostatistics and Analyses,
 Kamenice 3, 625 00 Brno,
 Czech Republic
 Email: {schwarz,
 janousova}@iba.muni.cz

Tomáš Kašpárek
 Masaryk University and University
 Hospital Brno, Department of
 Psychiatry, Jihlavska 20, Brno,
 Czech Republic
 Email: tkasparek@fnbrno.cz

Abstract—Computer-aided schizophrenia diagnosis is a difficult task that has been developing for last decades. Since traditional classifiers have not reached sufficient sensitivity and specificity, another possible way is combining the classifiers in ensembles. In this paper, we take advantage of random subspace ensemble method and combine it with multi-layer perceptron (MLP) and support vector machines (SVM). Our experiment employs voxel-based morphometry to extract the grey matter densities from 52 images of first-episode schizophrenia patients and 52 healthy controls. MLP and SVM are adapted on random feature vectors taken from predefined feature pool and the classification results are based on their voting. Random feature ensemble method improved prediction of schizophrenia when short input feature vector (100 features) was used, however the performance was comparable with single classifiers based on bigger input feature vector (1000 and 10000 features).

I. INTRODUCTION

SCHIZOPHRENIA (SZ) is a severe and chronic neurodevelopmental disorder with unknown etiology. Patient's response to the treatment is uncertain and early diagnosis could increase the probability of remission. Since nowadays the diagnostics is based on interview, self-report and psychiatrist's observation, there are efforts to develop a diagnostic tool that could support establishing diagnosis of the first episode of schizophrenia in a more objective way.

Involvement of modern imaging methods in the last decades has opened up new possibilities in brain research. These methods include for instance, magnetic resonance imaging (MRI), computed tomography or positron emission tomography. Especially MRI techniques offer good contrast and spatial resolution. Thus, morphological abnormalities and relations between brain structures and functions can be studied with the use of imaging data. Since the differences between schizophrenia patients and healthy controls (HC) have been already found with the use of manual segmentation of region of interests [1] or automated morphometry methods such as voxel-based morphometry [2] or deformation-based morphometry [3], many scientists have been recently trying to create computer-aided diagnostic tools based on neuroimaging data. The outcomes

of such tools have not reached sufficient sensitivity and specificity for implementation into the psychiatric clinical practice yet, and hence the demand still persists. The application of classification methods with self-adapting strategies known as machine learning is a challenging task in the schizophrenia research.

Artificial neural network (ANN) is a model inspired by how the brain works. Since the backpropagation algorithm [4] was invented as a technique for learning ANNs, they have been used widely in many applications and have achieved success at least in two areas of brain image processing: segmentation [5], [6] and classification [7]–[11]. Those results have shown that ANNs deserve attention of neuroimaging community investigating how to recognize mental diseases in imaging data.

Several authors have already tried to classify schizophrenia based on diffusion tensor data [7] using several types of neural networks: backpropagation neural networks, radial basis function networks, learning vector quantization neural networks and probabilistic neural networks. Other studies [8], [12] used backpropagation neural network on functional magnetic resonance imaging (fMRI) data or resting-state fMRI. Other papers discussed the use of ANN for classification of other brain diseases such as Alzheimer's disease [9], [10] or brain cancer [11] – based on structural MRI data.

The power of such single models can be further improved by the means of ensemble learning. This approach employs set of classifiers to determine the object's class by voting. To ensure a necessary assumption, which is the disparity among the classifiers, variability in the training process must be somehow acquired. Many methods have been invented for this purpose, such as random subspace ensemble [13], random forests [14], bagging [15], boosting (e.g. AdaBoost [16]), rotation forests [17] and others. The first one is explored in this paper in combination with ANNs. Several ensemble methods were applied to investigate neuroimaging problems. For instance, Yang et al. [18] classified schizophrenia using ensembles of support vector machines (SVM) trained by modified AdaBoost algorithm that besides boosting performed also simultaneous feature selection from

fMRI and single nucleotide polymorphism data. They reached an overall accuracy of 87% when both data sets were combined. Janousova et al. [19] based their ensemble for schizophrenia prediction on different image features extracted from MRI data (MR intensities, grey matter densities and local deformations) and on three various types of classifiers, and achieved the accuracy of 81.6%. Lebedev et al. [20] used random forests for Alzheimer's disease (AD) detection and achieved overall accuracy of 91%. Liu et al. [21] proposed local patch-based subspace ensemble method combined with a classifier based on sparse representation of data and improved the performance of classification performance of AD and mild cognitive impairment up to 3% compared to the use of a single classifier.

To the best of our knowledge, this is the first time that random subspace ensemble ANNs are used for schizophrenia classification based on the structural MRI data. The paper is organized as follows: Section II outlines the basics of ANNs and random subspace ensemble. The experiment and results of classification are summarized in section III. Section IV discuss the results and concludes the paper.

II. METHODS

A. Dataset and Image Processing

We used the same dataset as in [22]. It consisted of MRI data of 52 schizophrenia patients and 52 age- and sex-matched (only men) healthy control subjects without family history or personal history of axis I psychiatric conditions. Patients' and healthy controls' median age was 22.9 years (range 17-40 years) and 23.0 years (range 18.2-37.8 years) respectively. Images of all subjects were acquired on 1.5 T magnetic resonance imaging machine Siemens Symphony at University Hospital Brno, using a 3-D acquisition with IR/GR sequence, TR 1700 ms, TE 3.93 ms, TI 1100 ms, flip angle 15°, 160 slices, voxel size 1.17 × 0.48 × 0.48 mm, FOV 246 × 246 mm, and matrix size 512 × 512 voxels.

Gray matter tissue segments were obtained from all images after a correction of bias-field inhomogeneity, spatial normalization, segmentation, modulation and Gaussian smoothing. All the steps followed the pipeline of the optimized voxel-based morphometry [23]. The last step - smoothing the gray matter segments with an isotropic Gaussian kernel - spreads the information to the neighboring voxels and compensates the inexact nature of the spatial normalization step [24]. Furthermore it ensures more normally distributed data for further parametric statistical analyses used for feature selection [25]. After this image preprocessing, a binary mask of a brain normalized to the stereotactic space was used to erase the voxels representing extracerebral tissues.

B. Feature Pool Preparation

Before applying the classifiers, it was necessary to select only those features - gray matter density voxels - which represented the information useful to discriminate between the two classes, and conversely to exclude those voxels without any helpful information, and thus to improve signal-

to-noise ratio and create more accurate classifier. The feature selection step is also important to tackle the problem of the curse of dimensionality – well-known in neuroimaging community [26].

Voxel-wise two-sample t-tests were the instrument for feature selection. This method selected only those voxels which showed significant differences in gray matter density between the groups. Such a very naïve approach might be less prone to overfitting than selection the features with the highest discrimination power derived from correlation with the classification outcome. In addition, it is important that the whole volume of gray matter was explored with no arbitrarily predefined regions of interests, as morphological abnormalities in schizophrenics' brains have been uncovered with automated brain morphometry methods in many different brain regions [2].

C. Multi-layer Perceptron

Multi-layer perceptron (MLP) is the most traditional type of ANN. It maps relations between inputs and desired outputs. MLP consists of three or more layers formed by neurons - basic computational units - and are adapted in a supervised learning manner using the backpropagation algorithm. The information is passed through the layers in input-output direction. The equation of a neuron is:

$$y = \left(w_0 + \sum_{i=1}^n w_i x_i \right) \quad (1)$$

where y is the output, w_i are weights, x_i are inputs, n is the number of neuron inputs, σ is the activation function – hyperbolic tangent for hidden layers and linear functions in the output layer.

The output layer comprises two neurons; each represents one class, so the subject is classified according to the comparison of their values after the excitation. Furthermore, the softmax function is applied on these two neurons:

$$y_i = \frac{e^{\xi_i}}{\sum_{j=1}^n e^{\xi_j}}, \quad (2)$$

where n is the number of outputs, ξ_i and ξ_j are net activations of i -th and j -th output neurons. This function ensures that the outputs are non-negative and their sum is equal to 1. Such results can be interpreted as posterior probabilities [27].

In this project, the MLPs are trained by minimization of cross-entropy using the scaled conjugate gradient backpropagation algorithm that has been found to be fast in preliminary experimentation. Since neural networks have many parameters to be predefined - both for architecture creation and for adaption - it is difficult to set them to the optimal configuration. We kept the implicit learning parameters i.e. learning rate 0.01, maximal number of epochs 1000, minimum gradient 10^{-6} and regarding architecture and we used two-layer network with 10 hidden neurons, which achieved a good performance during experimentation. Since the weights were initialized

randomly, adaptation of the network was repeated 11 times and the classification of the testing subject was based on voting.

The Neural Network Toolbox for Matlab R2014b (The MathWorks, Inc.) was used here for all the described experiments.

D. Random Subspace Ensemble Principle

The datasets in neuropsychiatric research usually suffer from small sample size while their dimensionality is huge - often hundreds of thousands or even millions of voxels are used to describe each of subjects in the dataset. Hence, it is a hardly feasible task to avoid overfitting on a validation set by termination of neural network adaptation. Ensemble methods help to tackle overfitting and to improve the generalization ability.

In random subspace ensemble method, the variability is reached by a random selection of features from the set of all preselected features – a feature pool (FP). Each single classifier is adapted on one subset of feature which is less dimensional than the original feature space.

E. Validation and Evaluation

In order to report unbiased results, we used leave-one-out cross-validation strategy: one subject was left as a testing subject and the rest subjects composed a training set. This process was repeated *n* times, where *n* was the sample size. Since even only one testing subject withdrawn from the feature selection might have influenced the result, especially when the sample size was small, it was necessary to validate both features selection and classification.

For evaluation of the classification results, we used overall accuracy, sensitivity and specificity. The overall accuracy refers what proportion of subjects was classified correctly, whereas sensitivity and specificity say what the ability of diagnostic test is to reveal diseased and healthy person respectively.

III. EXPERIMENT AND RESULTS

The experiment was configured with many parameters – particularly size of the feature pool, number of chosen features and ensemble size. In order to achieve a reasonable computation time, we investigated only several predefined configurations. First, we defined a size of the feature pool as the 10000, 20000, 50000 and 100000 most significant voxels based on two-sample t-test criterion. This FP was used as the bag from which the features were chosen randomly.

The second parameter was a number of features used for the adaptation of the classifiers. We trained both MLP and SVM classifier on the different number of the most significant features - as shown in Fig. 1 - to find the optimum. The best performance was reached with the use of MLP on 1000 features and with the use of SVM on 100 features. Adding more features to the models did not reveal any trend in the classification performance - increasing or decreasing - so we later experimented with 3 options - 100, 1000 and 10000 - in order to explore both small and big dimensionality of the feature space. The MLP revealed

higher accuracies compared to the SVM in most configurations.

The last investigated parameter was the number of voting classifiers in ensemble, which was set to 31. This odd number ensures that the subject is always assigned to one of the classes – SZ or HC.

Since outcomes from the whole ensemble were available, i.e. 31 units, we could use for evaluation any combination of 1, 3, 5 etc. classifiers from this ensemble. Hence, we computed performance measures on all, but maximally 10000 combinations in order to gain nonrandom and computationally accessible outcomes. Furthermore, the subsets of features were chosen randomly, all experiments were repeated 10 times and the measures of the classification performances were averaged.

Figures 2-5 show the experimental results. In each figure, the outcomes computed on the different size of the feature pool are displayed.

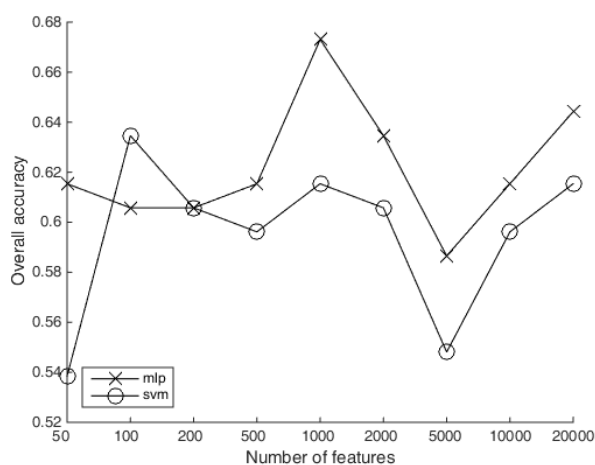


Fig. 1 The overall accuracy of the MLP and SVM dependent on the number of the most significant features selected with voxel-wise two-sample t-tests.

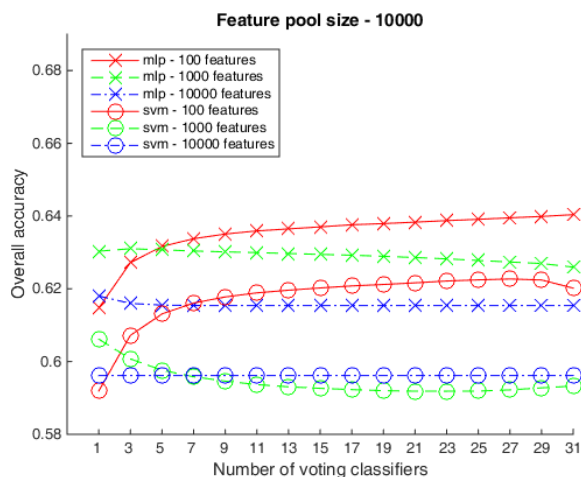


Fig. 2 Results of ensemble voting based on the feature pool with the size of 10000 selected features.

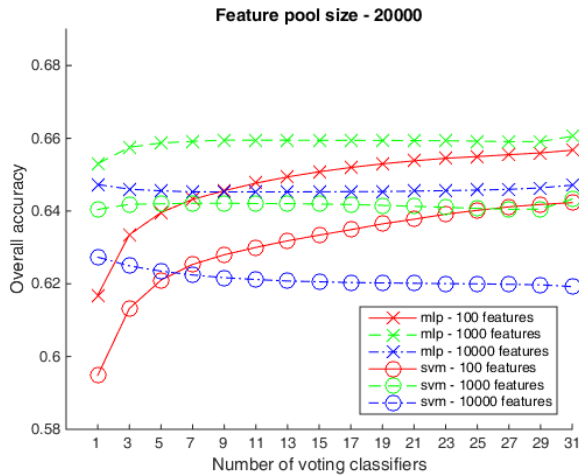


Fig. 3 Results of ensemble voting based on the feature pool with the size of 20000 selected features.

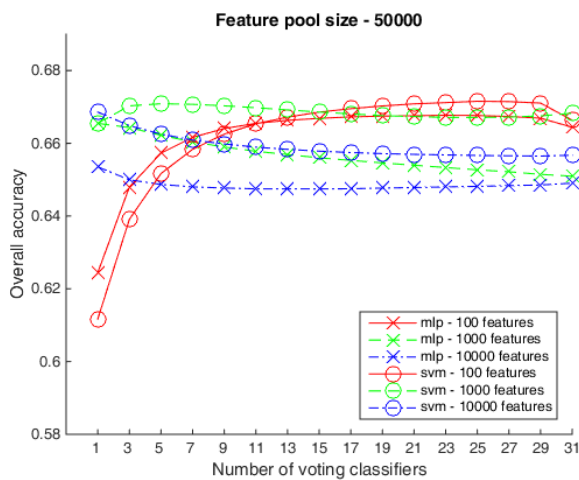


Fig. 4 Results of ensemble voting based on the feature pool with the size of 50000 selected features.

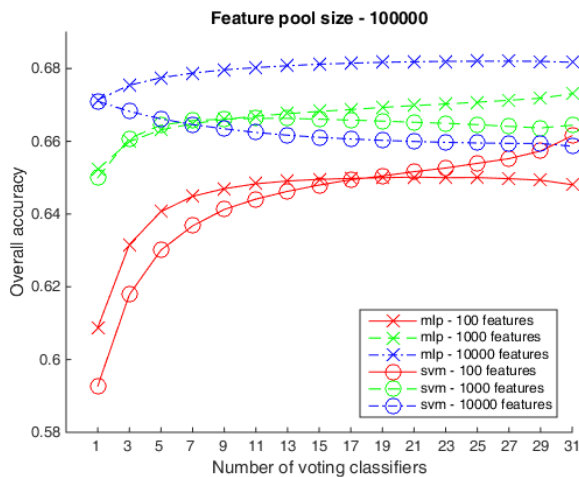


Fig. 5 Results of ensemble voting based on the feature pool with the size of 100000 selected features.

IV. DISCUSSION AND CONCLUSIONS

In this paper, we explored the random subset ensemble method for first-episode schizophrenia classification with

the use of multi-layer perceptron and performed a comparison to ensembles of SVM classifiers.

We selected the features with a naïve approach based on voxel-wise two-sampled t-tests. We believe that this approach may be less prone to overfitting than other more complex methods. More sophisticated and often computationally more time-consuming methods such as sequential forward/backward feature selection or multivariate extraction methods are left for our future research.

Since the neighboring voxels could have been correlated due to the smoothing in the image preprocessing phase, we assumed that higher amount of features in the feature pool enabled to capture the discriminative information in more parts of the brain, and therefore improved the classification performance. This improvement is observable when the Fig. 2 and 4 are compared.

Multi-layer perceptron was more effective when smaller feature pool was used. With the use of bigger feature pools, both classification methods yielded more similar outcomes.

Increasing the number of voting classifiers surprisingly improved the classification accuracy only in case of small number of input features. The increasing trend in the accuracy reached a level similar to as the models with more inputs and the increasing trend did not continue. The SVM with 100 features adapted on FP with 100000 features was improved by 6.88% and MLP with the same number of input features adapted on FP of size 50000 achieved improvement of 4.33% when compared to single variants of the classification methods.

The best outcomes were revealed with the use of MLP on 100000 input features and 29 voting classifiers that have the highest overall accuracy of 68% (sensitivity 67%, specificity 69%).

Finally, we compare our results of schizophrenia prediction to other studies dealing with MLP or ensemble learning. Jafri and Calhoun [8] achieved 75.6% with MLP based on fMRI data (38 SZ + 31 HC). Savio et al. [7] reached 100% on diffusion imaging data containing 20 male subjects using both neural networks and SVM. Yang et al. [18] also reached much higher overall accuracy 87% with SVM classifier ensemble based on AdaBoost, but besides small sample size (20 SZ + 20 HC) they admit the patients were chronic and under an antipsychotic medication. These studies could suffer from small sample size, since models tend to reach more variable outcomes and therefore it is easier to reach good (as well as poor) classification performance [28]. Since Janousova et al. [19] exceeded 81% with a similar number of subjects, we propose that an involvement of other classifiers and various feature extraction methods may be helpful in our framework too.

We conclude that the random feature ensemble method in combination with MLP and SVM improved prediction of schizophrenia from MRI data only in case of short feature vectors (100 features) – when compared to the use of single MLP or SVM classifiers. The classification accuracy achieved with the ensembles was not much different from

the accuracy of the single classifiers with feature vectors of higher dimensionality (1000 and 10000 features).

REFERENCES

- [1] J. Sun, J. J. Maller, L. Guo, and P. B. Fitzgerald, "Superior temporal gyrus volume change in schizophrenia: a review on region of interest volumetric studies," *Brain Res. Rev.*, vol. 61, no. 1, pp. 14–32, Jun. 2009. [Online]. Available: <http://dx.doi.org/10.1016/j.brainresrev.2009.03.004>
- [2] D. C. Glahn, A. R. Laird, I. Ellison-Wright, S. M. Thelen, J. L. Robinson, J. L. Lancaster, E. Bullmore, and P. T. Fox, "Meta-Analysis of Gray Matter Anomalies in Schizophrenia: Application of Anatomic Likelihood Estimation and Network Analysis," *Biol. Psychiatry*, vol. 64, no. 9, pp. 774–781, Nov. 2008. [Online]. Available: <http://dx.doi.org/10.1016/j.biopsych.2008.03.031>
- [3] C. Gaser, H.-P. Volz, S. Kiebel, S. Riehemann, and H. Sauer, "Detecting Structural Changes in Whole Brain Based on Nonlinear Deformations—Application to Schizophrenia Research," *NeuroImage*, vol. 10, no. 2, pp. 107–113, Aug. 1999. [Online]. Available: <http://dx.doi.org/10.1006/nimg.1999.0458>
- [4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.
- [5] J. Alirezaie, M. E. Jernigan, and C. Nahmias, "Neural network-based segmentation of magnetic resonance images of the brain," *IEEE Trans. Nucl. Sci.*, vol. 44, no. 2, pp. 194–198, Apr. 1997. [Online]. Available: <http://dx.doi.org/10.1109/23.568805>
- [6] Y. Li, Z. Li, and Z. Xue, "Segmenting MR Images Using Fully-Tuned Radial Basis Functions (RBF)," in *9th International Conference on Control, Automation, Robotics and Vision, 2006. ICARCV '06*, 2006, pp. 1–6. [Online]. Available: <http://dx.doi.org/10.1109/ICARCV.2006.345425>
- [7] C. J. Savio A, "Neural classifiers for schizophrenia diagnostic support on diffusion imaging data," *Neural Netw. World*, vol. 20, pp. 935–949, 2010.
- [8] M. J. Jafri and V. D. Calhoun, "Functional classification of schizophrenia using feed forward neural networks," *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. Suppl., pp. 6631–6634, 2006. [Online]. Available: <http://dx.doi.org/10.1109/IEMBS.2006.260906>
- [9] C. Huang, B. Yan, H. Jiang, and D. Wang, "Combining Voxel-based Morphometry with Artificial Neural Network Theory in the Application Research of Diagnosing Alzheimer's Disease," in *International Conference on BioMedical Engineering and Informatics, 2008. BMEI 2008*, 2008, vol. 1, pp. 250–254. [Online]. Available: <http://dx.doi.org/10.1109/BMEI.2008.245>
- [10] A. Savio, M. García-Sebastián, C. Hernández, M. Graña, and J. Villanúa, "Classification Results of Artificial Neural Networks for Alzheimer's Disease Detection," in *Intelligent Data Engineering and Automated Learning - IDEAL 2009*, E. Corchado and H. Yin, Eds. Springer Berlin Heidelberg, 2009, pp. 641–648. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04394-9_78
- [11] D. M. Joshi, N. K. Rana, and V. M. Misra, "Classification of Brain Cancer using Artificial Neural Network," 2010, pp. 112–116. [Online]. Available: <http://dx.doi.org/10.1109/ICECTECH.2010.5479975>
- [12] M. R. Arbabshirani, K. Kiehl, G. Pearlson, and V. D. Calhoun, "Classification of schizophrenia patients based on resting-state functional network connectivity," *Brain Imaging Methods*, vol. 7, p. 133, 2013. [Online]. Available: <http://dx.doi.org/10.3389/fnins.2013.00133>
- [13] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998. [Online]. Available: <http://dx.doi.org/10.1109/34.709601>
- [14] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: <http://dx.doi.org/10.1023/A:1010933404324>
- [15] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996. [Online]. Available: <http://dx.doi.org/10.1023/A:1018054314350>
- [16] Y. Freund and R. E. Schapire, *A Short Introduction to Boosting*. 1999.
- [17] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, Oct. 2006. [Online]. Available: <http://dx.doi.org/10.1109/TPAMI.2006.211>
- [18] H. Yang, J. Liu, J. Sui, G. Pearlson, and V. D. Calhoun, "A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia," *Front. Hum. Neurosci.*, vol. 4, Oct. 2010. [Online]. Available: <http://dx.doi.org/10.3389/fnhum.2010.00192>
- [19] E. Janousova, D. Schwarz, and T. Kasperek, "Combining various types of classifiers and features extracted from magnetic resonance imaging data in schizophrenia recognition," *Psychiatry Res. Neuroimaging*, vol. 232, no. 3, pp. 237–249, Jun. 2015. [Online]. Available: <http://dx.doi.org/10.1016/j.psychres.2015.03.004>
- [20] A. V. Lebedev, E. Westman, G. J. P. Van Westen, M. G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, and A. Simmons, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *NeuroImage Clin.*, vol. 6, pp. 115–125, 2014. [Online]. Available: <http://dx.doi.org/10.1016/j.nicl.2014.08.023>
- [21] M. Liu, D. Zhang, and D. Shen, "Ensemble Sparse Classification of Alzheimer's Disease," *Neuroimage*, vol. 60, no. 2, pp. 1106–1116, Apr. 2012. [Online]. Available: <http://dx.doi.org/10.1016/j.neuroimage.2012.01.055>
- [22] E. Janousova, D. Schwarz, G. Montana, and T. Kasperek, "Brain image classification based on automated morphometry and penalised linear discriminant analysis with resampling," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2015, pp. 263–268. [Online]. Available: <http://dx.doi.org/10.15439/2015F147>
- [23] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, Jul. 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.neuroimage.2005.02.018>
- [24] M. Kubicki, M. E. Shenton, D. F. Salisbury, Y. Hirayasu, K. Kasai, R. Kikinis, F. A. Jolesz, and R. W. McCarley, "Voxel-Based Morphometric Analysis of Gray Matter in First Episode Schizophrenia," *NeuroImage*, vol. 17, no. 4, pp. 1711–1719, Dec. 2002.
- [25] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *NeuroImage*, vol. 11, no. 6 Pt 1, pp. 805–821, Jun. 2000. [Online]. Available: <http://dx.doi.org/10.1006/nimg.2000.0582>
- [26] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–399, May 2011. [Online]. Available: <http://dx.doi.org/10.1016/j.neuroimage.2010.11.004>
- [27] R. O. Duda, *Pattern classification*, 2nd ed. New York: Wiley, 2001.
- [28] H. G. Schnack and R. S. Kahn, "Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters," *Neuroimaging Stimul.*, p. 50, 2016. [Online]. Available: <http://dx.doi.org/10.3389/fpsy.2016.00050>