Automatic Classification in Memory Clinic Patients and in Depressive Patients

Tolga Uslu, Lisa Miebach, Steffen Wolfsgruber, Michael Wagner, Klaus Fließbach, Rüdiger Gleim,Wahed Hemati, Alexander Henlein, Alexander Mehler
Goethe University, Frankfurt am Main,
Department for Neurodegenerative Diseases and Geriatric Psychiatry, University of Bonn, Bonn, Germany
German Center for Neurodegenerative Diseases, Bonn, Germany
{uslu, mehler, gleim}@em.uni-frankfurt.de
{lisa.miebach, steffen.wolfsgruber, michael.wagner,klaus.fliessbach}@ukbonn.de

Abstract

In the past decade the preclinical stage of *Alzheimer's Disease* (AD) has become a major research focus. *Subjective cognitive decline* (SCD) is gaining attention as an important risk factor of AD-pathology in early stages of *mild-cognitive impairment* (MCI), preclinical AD and depression. In this context, neuropsychological assessments aim at detecting sorts of subtle cognitive decline. Automatic classification may help increasing the expressiveness of such assessments by selecting high-risk subjects in research settings. In this paper, we explore the use of neuropsychological data and interview based data designed to detect AD-related SCD in different clinical samples to classify patients through the implementation of machine learning algorithms. The aim is to explore the classificatory expressiveness of features derived from this data. To this end, we experiment with a sample of 23 memory-clinic patients, 21 depressive patients and 21 healthy-older controls. We use several classifiers, including SVMs and neural networks, to classify these patients using the above mentioned data. We reach a successful classification based on neuropsychological data as well as on cognitive complaint categories. Our analysis indicates that a combination of these data should be preferred for classification, as we achieve an F-score above 90% in this case. We show that automatic classification using machine learning is a powerful approach that can be used to improve neuropsychological assessment.

Keywords: early diagnostics, disease classification, feature selection, Alzheimers Disease, neuropsychology

1. Introduction

According to the world Alzheimer report, over 46 million people are estimated to have dementia. This number is expected to rise (Prince et al., 2015). Early detection and accurate diagnostic in preclinical stages is therefore of paramount importance. As an indicator of the earliest clinical stage of Alzheimers Disease (AD) subjective cognitive decline (SCD), defined as the individual's concerns related to cognitive functioning, is gaining interest in different settings (Jessen et al., 2014). With the growing interest in early diagnosis and early detection, SCD has been proposed as an established risk factor for AD, increased risk of future cognitive decline (Koppara et al., 2015) and abnormal AD biomarkers (Amariglio et al., 2012; Chetelat et al., 2010; Wolfsgruber et al., 2015; Buckley et al., 2017). However, in older community based samples the prevalence of memory concerns varies from 25-50% (Jonker et al., 2000) which made it difficult to distinguish AD-related cognitive complaints from those related to normal aging. Furthermore, subjective cognitive decline (SCD) is reported in the context of depression (Balash et al., 2013) and has been positively associated with SCD in different samples (Buckley et al., 2013; Benito-León et al., 2010). Some researcher therefore argued that SCD is mainly driven by depressive symptomatology than being an indicator of an underlying AD-pathology. Current investigations tries to refine the assessment of SCD with the aim to find AD-like complaints and those which may be more representative of a mood disorder or of aging in general (Molinuevo et al., 2016; Rabin et al., 2015). In line with the problematic assessment of SCD, some common-used neuropsychological screening tests such as the *Mini-Mental State Examination* (MMSE) are not sensitive enough for a reliable detection of subtle impairments presented in patients with *mild cognitive impairment* (MCI). Even when some results suggest specific types of neuropsychological deficits associated with *Major depressive Disorders* (MDD), it is still challenging for clinicians to differentiate subjective complaints as a result of a depressive symptomatology from cognitive complaints in the context of preclinical and prodromal AD (Zihl et al., 2010). In memory-clinic settings, early detection of AD is time consuming and require multiple cost intensive information (e.g. neuropsychological testing including subjective concerns and objective impairment, detailed medical history and neurological examination) as well as clinicians with a certain level of expertise. Current assessments of subjective cognitive decline are unable to capture all aspects of SCD specific for preclinical AD and could potentially confound results in the SCD field. Recently, studies started to compare specific aspects of cognitive complaints in different samples using qualitative interview based approaches (Buckley et al., 2015; Miebach et al., 2017; Miebach et al., 2018)

In conclusion, there is large room for improvement regarding the quantitative assessment of SCD and subtle cognitive decline which pose a major task for further research (Jessen et al., 2014). Automatic classification and machine learning might help detecting specific assessment strategies for preclinical AD and the refinement of neuropsychological test batteries.

We generated various neuropsychological and clinical parameters from patient conversations and examinations. To allow automatic classification using this data, we used multiple types of classifiers (SVM, neural networks) to make a diagnosis. In some cases we even managed to get a classification reaching an F-score of more than 90%.

In any event, it is very time-consuming to generate the underlying medical data. Therefore, it is of utmost importance to generate only those data that is required to pro-

duce a good classification. To find out this data, we evaluated different approaches. On the one hand, we used a genetic search over the underlying feature space to find out which subset of features leads to better results. On the other hand, we calculated distance correlation to detect dependencies between pairs of features. We discovered that in some cases, less than 50% of the features of the underlying medical study suffice to generate the best performing classification.

2. Related Work

Machine learning techniques are becoming more and more popular in clinical research and are an established technique in MRT studies (Bede, 2017). Recent studies start from optimizing neuropsychological assessment for cognitive, behavioral and functional impairment using machine learning (Battista et al., 2017). However, studies using automatic classification to distinguish AD from non-AD patients did not focus on earlier preclinical or early MCI stages (Gurevich et al., 2017). Further, modern machine learning techniques have up to now only very rarely been used for the differential diagnosis of cognitive complaints based on the results of interview data. Mehler et al. (2016), for example, automatically analyzed physician-patient talks for differentiating patients suffering from epilepsies or dissociative disorders. This was done by means of the *text2voronoi* algorithm, which is also used in this paper. Regarding the assessment of SCD, (Miebach et al., 2017) were able to confirm several qualitative complaint categories proposed by (Buckley et al., 2015) which are specific for memory-clinic and depressive patients. This suggests that the subjective experience of cognitive decline can be captured by means of a set of interview questions and categories and therefore could be useful for clinicians to detect individuals at high-risk for AD. Investigations of MCI patients self-awareness and experience of their diagnosis have revealed that qualitative approaches may well lead to a more in-depth view than quantitative measurements (Lingler et al., 2006; Roberts and Clare, 2013). However, a qualitative approach is more time consuming than a quantitative one making the diagnostic process more cost intensive. With the gaining interest in an improved detection rate of AD-pathology with less time and cost intensive screening tools, clinicians have the unique opportunity to take advantage of automated classification techniques. This exploratory example of machine learning combined neuropsychological data for the assessment of cognitive impairment and qualitative extracted interview-based features for cognitive complaints in memory-clinic-patients, depressive patients and in healthy controls.

3. Models

In the present study, we experiment with several classification models to be independent of the classifier and to assess the significance of features while being less dependent on these classifiers. As input, the classifiers are fed with neuropsychologically and clinically determined feature values. The neuropsychological part of our study includes a test battery for assessing cognitive performance and depressive symptoms. The clinically determined values are ratings based on qualitative interviews designed to capture aspects of subjective cognitive complaints in the context of preclinical dementia. In contrast to the neuropsychological data set these values are based on expert ratings instead of self-ratings or performance measures. The different group status (memory-clinic-patients, depressive patients, healthy controls) were set as output.

Since we only have a limited amount of data, we carried out a leave-one-out cross-validation for each classifier being tested. This makes sense since each patient is referred to individually for classification. With other data splitting methods, the risk of overfitting is too high (achieving good results on a given split, while performing bad on another one).

3.1. SVM

As a baseline for the experiments we trained a *Support Vector Machine* (SVM) and used it for classification. This is done by means of the SVM-light (Joachims, 1998) implementation using the radial basis function (RBF) kernel. To find optimal parameters for training, we carried out a parameter study on the gamma and the cost parameter. For the cost parameter we examined values between 0.01 and 0.000001; for the gamma parameter we considered values in the range of 1 and 1000000.

3.2. Neural Network

To carry out the same experiments using modern classification methods, neural network-based methods were indispensable. To this end, we used the framework Keras (Chollet and others, 2015). More specifically, we trained a feed-forward network to get a classifier of medical data. Here again, we conducted a parameter study to find the best performing parameters in each experiment. The following parameters were evaluated:

- optimizer: [adam, adamax, rmsprop]
- dropouts: [0.25, 0.5, 0.75]
- layersize: [50, 100, 200, 500]
- layersize2: [0, 50, 100, 200]

We achieved the best results with a dropout of 0.25, *adam* (Kingma and Ba, 2014) as optimizer and two hidden layers.

3.3. Systematic Feature Evaluation for SVM

We examine the impact of feature selection on the F-Measure. While some features may consistently contribute to good classification results, others may reduce performance. That is, we expect that using all available features will most likely not yield the best F-Measure. Since a systematic evaluation of all $2^{138} - 1$ feature combinations is impossible, we apply several approaches to determine local optimal values and to examine the overall robustness of the feature set. If not stated otherwise each evaluation of a given feature set includes a parameter study regarding the optimal *gamma* and *cost* value for the SVM. Here again, our studies are based on SVM-light (Joachims, 1998).

We start with performing a genetic search for optimal feature selection. Genetic algorithms have successfully been

used for feature selection (Li et al., 2005). In our case, a population of n variants, which have been initialized randomly, are evaluated, ranked and flipped (bitwise) over t turns. In each turn, the best ranking variants are kept and mutated to generate additional variants while worst performing instances are dropped. In this way, a hill-climbing algorithm is implemented that approaches local maxima of better performing subsets of features.

3.3.1. Top-down and bottom-up search

In order to examine the overall robustness of the feature set we gradually remove features from the entire set (top-down).

In addition we explore the effect of gradually increasing the number of features starting from an empty set (bottom-up). At each step, the feature that maximizes the performance of the remaining set is added or removed, resulting in $\frac{n^2+n}{2}$ computations. Whenever multiple variants achieve the same top value we chose one of them randomly.

Applying this methodology to feature reduction is an important step, as it not only improves the classification results but also helps reducing the computation time in further analyses.

3.4. text2voronoi

Mehler et al. (2016) have developed a new classification method which visualizes input texts and then uses the visual representation of these texts to drive the classification. The advantage of this method is that one gets a visual depiction of the underlying text that can be used by analogy to MRI scans. Instead on working on the content words of a text, *text2voronoi* is mainly working on distributions of grammatical features of words in this text. In this way, it allows for completely abstracting from text content. This is indispensable when dealing with rather short talks of doctor and patients which, though describing the same disease, may select words of a completely unrestricted semantic universe. Using grammatical information, embeddings are produced by means of word2vec (Mikolov et al., 2013). Then, a Voronoi tessellation is calculated on this data to map texts onto 2- or 3D spaces. Finally, the resulting depictions are used explored to drive the classification.

3.5. fastText

We additionally experimented with fastText (Joulin et al., 2016), an efficient text classifier, to compare it with *text2voronoi*. fastText is based on a feedforward neural network with only one hidden layer. Joulin et al. (2016) show that fastText compares with state-of-the art classifiers while being faster than its competitors.

4. Experiment

4.1. Sample description

The total sample of this study includes n=65 older subjects (mean age=70.03 years; 52.3% female). All participants were above the age of 55 and had sufficient ability to speak German. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study was approved by the local ethical committees of the University of Bonn, and informed written consent was obtained from all subjects.

Memory-clinic patients (MC)(n=23) were referred by their general practitioners to the Clinical Treatment and Reseach Center for Neurodegenerative Disorders (KBFZ), Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn for a diagnostic work up of cognitive functioning. Diagnosis of AD Dementia or MCI was made according to the core clinical criteria of the NIA-AA (Albert et al., 2011)(McKhann et al., 1984). The diagnostic procedure included a cognitive assessment, detailed medical history, and a neurological examination. Of the total sample 15 fulfilled the core clinical criteria of mild cognitive impairment (MCI) according to the NIA-AA criteria (performance under 1.5 SD below age, gender, education adjusted norms)(Albert et al., 2011). The remaining 8 patients only had subjective concerns without objective impairment, and were classified as patients with Subjective Cognitive Decline (SCD).

Major depressive Patients (MDD) (n=21) were recruited from the Clinic of Psychiatry and Psychotherapy, University Hospital Bonn. All patients fulfilled a diagnosis of a unipolar, major depressive disorder according to ICD-10 criteria (Organization, 1993).

The Healthy control group (HC) (n=21) was recruited from a scope of a normative study of the German Center for Neurodegenerative Diseases (DZNE) Bonn that evaluated neuropsychological performance of healthy older individuals. They were excluded from the participation when they (1) were concerned about mental abilities or memory (2) had been in psychological, psychiatric or neurological treatment within the last 6 months (3) had any severe or chronic disease (e.g. diabetes or MS) (4) had experienced head injury with a loss of consciousness, (5) had a neurological disease (e.g. AD or Parkinson) (6) or had a relative with a first-degree relative with a documented diagnosis of neurodegenerative disease in their family history.

4.1.1. Clinical Rating of cognitive complaints

The Clinical rating was made based on a semi-structured interview designed to capture all complaint categories proposed by (Buckley et al., 2015). The Interview similar to a clinical routine interview, started with an open question asking whether the patient had noticed *"any changes in memory or thinking during the last years*" followed by detailed questions about the complaint itself. The interview procedure followed a semi-structured format and lasted between 8 and 31 min. Each interview had an unstructured beginning, which allowed patients to determine the initial focus of the conversation. If cognitive changes were reported, the participants were asked to give an example of their everyday life. Then the patient was asked whether he/she has noticed further cognitive problems followed by the request to give an everyday example. This process was repeated until the participant did not mention further complaints. He/she was then asked to name the most concerning symptom which was selected for further detailed questioning. If the participant reported another concerning symptom, we repeated the detailed questions

about the complaint itself. Therefore, 58% of the sample named two concerning symptoms. All Interviews were digitally recorded and later transcribed verbatim by the interviewer. Data for analyses presented in this manuscript were derived from the ratings of a single clinical psychologist (LM) who also conducted all the interviews. To capture all aspects of cognitive complaints, the clinical rating in this study was based on glossary of cognitive complaints based on a combination of the cognitive complaint categories proposed by (Buckley et al., 2015) and the complaint themes proposed by (Miebach et al., 2018). The glossary contains the following categories: *Increasing frequency, Sense of predomination and growing concern, Situational lapses, Relative absence of spatio-temporal contextualisation, burdensome coping strategies, Dismissive attitude, attentional fluctuation/vagueness, Impact on affect, Progression, an over-endorsed complaint, dependency, affective influence on memory, distractible speech, general complaints about increasing memory problems, difficulties in Action monitoring, difficulties in initiating actions, deceleration, slowing of cognitive processing speed, nonspecific overwork, forgetfulness, short-term memory problems, content memory problems, blank mind, loss-of-control experience, derealisation, formal though disorder, prospective memory, planning, learning, cognitive flexibility, increased distractibility, concentration difficulties, word finding difficulties, memory for names, dyscalculia, visual-spatial-disorientation, general decline, no changes in cognitive functions.* The categories were extracted from the interview material using inductive qualitative approaches. The complaint categories based on (Buckley et al., 2015) were related to the grounded theory (Strauss and Corbin, 1997) whereas the complaint themes extracted by (Miebach et al., 2018) were based on the interpretative phenomenological analysis (IPA) (Smith et al., 2009). Therefore the presented deductive rating of cognitive concerns is based on two different phenomenological approaches which allows to capture highly nuanced and contextualized aspects of subjective experiences (Smith et al., 2009). The Interview procedure and categorization system are described in detail in (Miebach et al., 2017; Miebach et al., 2018). For the coding process, we used a deductive category assignment approach similar to qualitative content analysis (Mayring, 2014). Participant's responses were coded using a binary coding system (i.e. 0=theme absent; 1=theme present).

4.1.2. Neuropsychological assessment

The Neuropsychological assessment included a set of different clinical measurements for global memory and cognitive performance specifically designed for early diagnosis of AD dementia. The test battery included the Free and Cued Selective Reminding Test (FCSRT) (Ivnik et al., 1997) and the German version of the neuropsychological test battery of the Consortium to Establish a Registry for Alzheimers disease (CERAD-plus;(Morris et al., 1989)) with various sub-tests (e.g. verbal fluency, Boston Naming Test, Mini Mental State Exam, Word List learning, Constructional praxis, Word List recall, word list recognition, constructional praxis recall, TMT-A, TMT-B, the symbol digit modalities test (SDMT)(Smith, 1982)). Depressive symptoms were assessed with the 15-item version of the Geriatric Depression Scale (GDS; (Yesavage et al., 1983)) and the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2010)

4.1.3. Group characteristics and demographical differences

Analysis for the group differences were performed using IBM SPSS Version 22 (Corp, 2013). Group differences were observed for age, education, interview duration, GDS and PHQ-9 scores. Memory-clinic patients were slightly older (M=72.91 yr) compared to MDD (M=69.43 yr) and the interview duration was significantly longer (M=18.41 min) in comparison with HC (M=14.32 min) and the MDD-Group (M=12.07 min). HC were younger, performed significantly better on the MMSE (M=29) and exhibited lower levels of depressive symptomatology (GDS; M=0.62) compared to MDD and the Memory-clinic patients. The depressive group exhibited elevated levels of depressive symptomatology, significantly above the GDS cut-off for depression (M=7.00) and the PHQ-9 cut-off for moderate depression (M=10.89). Depressive patients also had significantly fewer years of education (M=12.57) compared to HC (M=15.10) and Memory-clinic patients (M=15.50).

4.2. Classification

We have used the models from section 3. to classify the patients. In doing so, we classify on the textual data and on the clinical and neuropsychologically generated data. Models 3.1. to 3.3.1. are designed for the classification with the self-generated data, while the models 3.4. and 3.5. are designed for textual classification.

4.2.1. Clinical and neuropsychological feature classification

In this experiment, we used the clinical ratings of subjective cognitive complaints based on the qualitative interview as one feature set. We also used the neuropsychological test results including data about objective cognitive performance and measurements of depression as another feature set. First of all, we used both feature sets independently for classification. We then combined both sets of features and used the combination for classification. Table 1 shows the results of the different classifiers applied to the 3 feature sets. We discovered that the combined features are always more successful than both feature sets on their own.

Model	Neuropsych.	Clinical	Combined
3.1.	0,747	0,706	0,794
3.2.	0,754	0,723	0,800
3.3.	0,870	0,821	0,881
3.3.1.	0,933	0,928	0,949

Table 1: F-scores of the classifiers with regard to the different data sets.

4.2.2. Patient talks classification

In this experiment, we analyze the texts of the cognitive complaint interviews and use them for classification (leave one out cross validation). We only used the text content of

the patients and removed the doctor's text data from the interview protocols because the doctor asks all patients similar questions, which would have a negative effect on the classification. Table 2 shows the results of the 2 methods we used for classification. It can be seen that the baseline classifier *fastText* can hardly classify the texts. However, if the texts are abstracted, as it is the case with *text2voronoi*, an improvement is achieved.

Model	F-score
3.4.	0,520
3.5.	0,340

Table 2: Results of the textual classification experiment of the 3 patient groups (MC/MDD/HC)

4.3. Feature analysis

Now that we have applied different classifiers in different experiments, we want to find out which of the used features were actually needed. To find out, we have used the following approaches.

4.3.1. Genetic feature search

As explained in chapter 3.3., we have also carried out a genetic search of the parameters to find the smallest possible subset, which provides the best results. We found out that only a fraction of the features are required to perform a good classification.

Experiment	Subset
Neuropsych.	47,30%
Clinical	40,63%
Combined	40,58%

Table 3: Subset analysis of the features using model 3.3..

In addition to the genetic search, we have also carried out two other approaches, as described in Section 3.3.. Figure 1 shows the process of this analysis. Again, it is obvious that few features are enough to get the best results. We achieve the best score with only 64 features.

4.3.2. Decision tree

A good way to analyze the features is to use Decision Trees, as it follows simple and comprehensible heuristics. The graphic representation as a tree diagram also illustrates hierarchically consecutive decisions. We have used the Python package *sklearn* (Pedregosa et al., 2011) to perform these analyses. In our best performing experiment, we have the following patient distribution:

- $[21, 23, 21]$ - (Control patients, Memory-Clinic patients, Depressive patients)

Figure 2 shows that feature 41 (SDMT - neuropsychological score developed to identify individuals with neurological impairment) is at the top of the tree. If the value of SDMT is less than or equal to 44,5, the patient distribution is divided into:

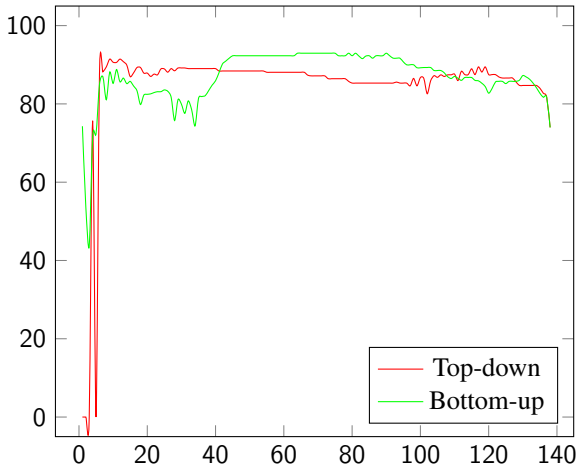

Figure 1: F-Scores based on the number of features in the example of the combined experiment (see Section 4.2.1.).

- $[0, 20, 18]$ - Only memory-clinic or depressive patients.
- $[21, 3, 3]$ - Mostly healthy controls.

Thus, we divide all control patients into a separate group. The next important parameter is feature 45 (GDS - neuropsychological measurement for depression)(Yesavage et al., 1983). This divides the group of diseases ($[0, 20, 18]$) into the following patient distribution:

- $[0, 17, 4]$ - Mostly Memory-clinic patients
- $[0, 3, 14]$ - Mostly depressive patients

This means that we could group 52 (14+17+21) patients correctly with these 2 features alone, but 13 (3+4+3+3) wrong. You can also see that these features belong to the neuropsychological features. This also makes sense, as these values also lead to better classifications (see example 4.2.1.). The further down the tree is examined at, the more precise the distributions will be. However, given the number of features (138) and the small amount of patients (65), this is rather overfitting.

4.3.3. Distance correlation

To measure the interdependence between the features as described in Section 4.1. we calculated distance correlation between pairs of features. For this we used the R package *energy* (Rizzo and Szekely, 2017) utilized by *TextImager* (Uslu et al., 2017). Interdependent features are an indicator for redundant data. These redundant features are less helpful for classification. Figure 3 shows the heatmap of the pairwise dependencies. Each cell represents the distance correlation of the features X and Y, with the green shading indicating the dependency (darker = more dependent). The diagonal is green which indicates that every feature is correlated to itself. The green squares also provide important information. The first dependency square at the top left in Figure 3, for example, contains only neuropsychological features based on the MMSE. The mini-mental state test (MMSE) is a brief screening tool for Alzheimer dementia and impairment in global cognition (Folstein et al., 1975).

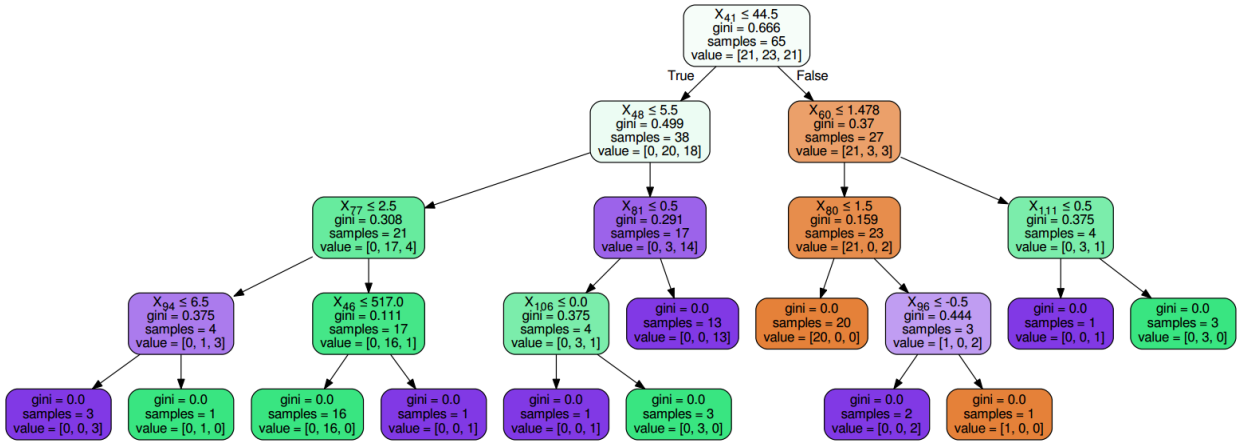

Figure 2: Result of the decision tree based on the combined example of experiment 4.2.1..

The MMSE test includes items assessing orientation, word recall and registration, attention and calculation, and language and visuospatial abilities. As a logical consequence we observed a high dependency between the different subscores of the MMSE.

As mentioned above, 58% of the sample named more than one cognitive complaint. As a result, the categories were coded for a second time. A high dependency between these features is therefore a consequence of the interview procedure. These dependencies can be seen in Figure 3 by the large green square at the bottom right.

Figure 3: Visual depiction of pairwise dependence of the features.

5. Discussion

The present study is the first to combine a qualitative text-analytical approach for cognitive complaints with an automatic classification system for three different diagnostic groups. Recently, only a few studies have explored the use of automated learning methods within the neuropsychological literature. In this proof of principle study we used a machine learning approach based on neuropsychological and interview generated cognitive complaint categories for the classification of memory-clinic patients, depressive patients and normal healthy older adults.

We aimed to replicate the diagnostic value of the recently proposed complaint categories using an automatic classification method instead of current statistical methods used in clinical research (Buckley et al., 2015; Miebach et al., 2017; Miebach et al., 2018). Cognitive complaints were elicited with a semistructured interview comparable with a typical clinical routine interview.

The current study results revealed that machine learning techniques can accurately classifying patients measured via neuropsychological test battery and via clinical rating of cognitive complaints. We found that the classification with self-generated characteristics extracted by a qualitative approach works much better than with the recorded texts in the patient conversations.

This result makes sense because patients talk about many different topics in the diagnostic interview and the content of these texts is not reliable for determining a disease. Therefore the interpretation of cognitive complaints relies on expertise of some kind which is not ideal for a wide distribution across studies.

In the case of the second experiment, the neuropsychological data outperformed the clinical ratings based on interview data. This could be explained by the heterogeneous sample including patients with mild cognitive impairment as well as patients with only subjective cognitive decline and depression. In line with current literature, the combination of neuropsychology and the clinical rating reached the best diagnostic accuracy (Molinuevo et al., 2016). A replication in a larger sample with focus on the complaint categories is needed to extract features which are truly relevant for AD-pathology.

Given the present results, we believe it is much more likely that measure incorporates both qualitative text based and quantitative neuropsychological methods will be able to

identify the preclinical AD profile. Recent studies used composite scores calculated based on z-transformed subscales of different SCD assessments to predict the taupathology in the enthorinal cortex of healthy older adults (Buckley et al., 2017).

However, in the case of textual classification, an improvement is achieved when the text is transformed into a more abstract model (*text2voronoi*). The experiments also show that the neuronal network-based approaches are usually somewhat better than the SVMs. However, the best solutions can be found with GeneticSVM and even only a subset of all features. As a result, we found out that a few features are enough to get a good classification. However, these features (SDMT; GDS) are established clinical screening tools for the measurement of memory impairment and depressive symptomatology (Yesavage et al., 1983; Smith, 1982). A feature analysis only including the cognitive complaint categories should be an important next step with a higher clinical impact in the field of AD research. We analyzed them and found out that there are some dependencies among the features. There is a need of alternative ways for the operationalization and the diagnosis of AD-relevant cognitive complaints. Using a semistructured interview based on qualitative categories seems to be promising regarding the clinical evaluation of memory complaints in non-demented elderly. Further improvement of the complaint glossary and the rating scale is needed for the detection of preclinical AD. Therefore machine learning approaches could be promising for reducing and refining neurospychological assessments. This information can save a lot of work, since the dependent features barely improve the classification.

6. Conclusion

In this paper we have used different classifiers in various patient diagnosis experiments. We have shown that a good classification can be achieved by using cognitive complaint categories based on clinical interview and neuropsychological data from standardized test batteries. We found that the combination of these data sets leads to the best results with an F-score of 80,00%. In addition, we have applied a number of different approaches to find the optimal subset of features that provide the best classification. In this case we even achieve an F-score of 94,87%. However, classification at text level is not yet particularly successful. In future work we aim at studying different abstractions of texts (as provided, for example, by *text2voronoi*) in order to detect expressive linguistic features that allow for automatically assessing the diseases under consideration.

7. Bibliographical References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., et al. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):270–279.

Amariglio, R. E., Becker, J. A., Carmasin, J., Wadsworth, L. P., Lorius, N., Sullivan, C., Maye, J. E., Gidicsin, C., Pepin, L. C., and Sperling, R. A. (2012). Subjective cognitive complaints and amyloid burden in cognitively normal older individuals. *Neuropsychologia*, 50(12):2880–2886.

Balash, Y., Mordechovich, M., Shabtai, H., Giladi, N., Gurevich, T., and Korczyn, A. D. (2013). Subjective memory complaints in elders: depression, anxiety, or cognitive decline? *Acta Neurologica Scandinavica*, 127(5):344–350.

Battista, P., Salvatore, C., and Castiglioni, I. (2017). Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study. *Behavioural neurology*, 2017.

Bede, P. (2017). From qualitative radiological cues to machine learning: Mri-based diagnosis in neurodegeneration.

Benito-León, J., Mitchell, A. J., Vega, S., and Bermejo-Pareja, F. (2010). A population-based study of cognitive function in older people with subjective memory complaints. *Journal of Alzheimer's Disease*, 22(1):159–170.

Buckley, R., Saling, M. M., Ames, D., Rowe, C. C., Lautenschlager, N. T., Macaulay, S. L., Martins, R. N., Masters, C. L., O'Meara, T., Savage, G., et al. (2013). Factors affecting subjective memory complaints in the aibl aging study: biomarkers, memory, affect, and age. *International Psychogeriatrics*, 25(8):1307–1315.

Buckley, R. F., Ellis, K. A., Ames, D., Rowe, C. C., Lautenschlager, N. T., Maruff, P., Villemagne, V. L., Macaulay, S. L., Szoeke, C., and Martins, R. N. (2015). Phenomenological characterization of memory complaints in preclinical and prodromal alzheimer's disease. *Neuropsychology*, 29(4):571.

Buckley, R. F., Hanseeuw, B., Schultz, A. P., Vannini, P., Aghjayan, S. L., Properzi, M. J., Jackson, J. D., Mormino, E. C., Rentz, D. M., Sperling, R. A., et al. (2017). Region-specific association of subjective cognitive decline with tauopathy independent of global β-amyloid burden. *JAMA neurology*, 74(12):1455–1463.

Chetelat, G., Villemagne, V. L., Bourgeat, P., Pike, K. E., Jones, G., Ames, D., Ellis, K. A., Szoeke, C., Martins, R. N., and O'Keefe, G. J. (2010). Relationship between atrophy and b–amyloid deposition in alzheimer disease. *Annals of neurology*, 67(3):317–324.

Chollet, F. et al. (2015). Keras. `https://github.com/keras-team/keras`.

Corp, I. (2013). Ibm spss statistics for windows, version 22.0. *Armonk, NY: IBM Corp*.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

Gurevich, P., Stuke, H., Kastrup, A., Stuke, H., and Hildebrandt, H. (2017). Neuropsychological testing and machine learning distinguish alzheimer's disease from other causes for cognitive impairment. *Frontiers in aging neuroscience*, 9.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Tangalos, E. G.,

Kokmen, E., and Petersen, R. C. (1997). Free and cued selective reminding test: Moans norms. *Journal of Clinical and Experimental Neuropsychology*, 19(5):676–691.

Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K. A., Van Der Flier, W. M., et al. (2014). A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease. *Alzheimer's & dementia*, 10(6):844–852.

Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Jonker, C., Geerlings, M. I., and Schmand, B. (2000). Are memory complaints predictive for dementia? a review of clinical and population-based studies. *International journal of geriatric psychiatry*, 15(11):983–991.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koppara, A., Wagner, M., Lange, C., Ernst, A., Wiese, B., Koenig, H.-H., Brettschneider, C., Riedel-Heller, S., Luppa, M., Weyerer, S., et al. (2015). Cognitive performance before and after the onset of subjective cognitive decline in old age. *Alzheimers & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(2):194–205.

Kroenke, K., Spitzer, R. L., Williams, J. B., and Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General hospital psychiatry*, 32(4):345–359.

Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., Wang, Q., Topol, E. J., Wang, Q., and Rao, S. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 85(1):16 – 23.

Lingler, J. H., Nightingale, M. C., Erlen, J. A., Kane, A. L., Reynolds III, C. F., Schulz, R., and DeKosky, S. T. (2006). Making sense of mild cognitive impairment: a qualitative exploration of the patient's experience. *The Gerontologist*, 46(6):791–800.

Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of alzheimer's disease report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–939.

Mehler, A., Uslu, T., and Hemati, W. (2016). Text2voronoi: An image-driven approach to differential diagnosis. In *VL@ ACL*.

Miebach, L., Wolfsgruber, S., Frommann, I., Buckley, R., and Wagner, M. (2017). Different cognitive complaint profiles in memory clinic and depressive patients. *The American Journal of Geriatric Psychiatry*.

Miebach, L., Wolfsgruber, S., Frommann, I., Fließbach, K., Jessen, F., Buckley, R., and Wagner, M. (2018). Cognitive complaints in memory clinic patients and in depressive patients: An interpretative phenomenological analysis. *The Gerontologist*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Molinuevo, J. L., Rabin, L. A., Amariglio, R., Buckley, R., Dubois, B., Ellis, K. A., Ewers, M., Hampel, H., Klöppel, S., and Rami, L. (2016). Implementation of subjective cognitive decline criteria in research studies. *Alzheimer's & Dementia*.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., Mellits, E. D., and Clark, C. (1989). The consortium to establish a registry for alzheimer's disease (cerad): I. clinical and neuropsychological assessment of alzheimer's disease. *Neurology*.

Organization, W. H. (1993). *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, volume 2. World Health Organization.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prince, M., Wimo, A., Guerchet, M., Ali, G., Wu, Y., and Prina, M. (2015). Alzheimer's disease international (2015). world alzheimer report 2015: The global impact of dementia: An analysis of prevalence, incidence, cost and trends. *Alzheimer's Disease International, London*.

Rabin, L. A., Smart, C. M., Crane, P. K., Amariglio, R. E., Berman, L. M., Boada, M., Buckley, R. F., Chételat, G., Dubois, B., and Ellis, K. A. (2015). Subjective cognitive decline in older adults: An overview of self-report measures used across 19 international research studies. *Journal of Alzheimer's Disease*, (Preprint):1–25.

Rizzo, M. L. and Szekely, G. J. (2017). Package 'energy'.

Roberts, J. L. and Clare, L. (2013). Meta-representational awareness in mild cognitive impairment: An interpretative phenomenological analysis. *Aging & Mental Health*, 17(3):300–309.

Smith, J., Flowers, P., and Larkin, M. (2009). Interpretative phoneomological analysis: theory, method and research.

Smith, A. (1982). Symbol digit modalities test (sdmt) manual (revised) western psychological services. *Los Angeles*.

Strauss, A. and Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Uslu, T., Hemati, W., Mehler, A., and Baumartz, D. (2017). Textimager as a generic interface to r. *EACL 2017*, page 17.

Wolfsgruber, S., Jessen, F., Koppara, A., Kleineidam, L., Schmidtke, K., FrÃ¶hlich, L., Kurz, A., Schulz, S., Hampel, H., and Heuser, I. (2015). Subjective cognitive decline is related to csf biomarkers of ad in patients with mci. *Neurology*, 84(12):1261–1268.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang,

V., Adey, M., and Leirer, V. O. (1983). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1):37–49.

Zihl, J., Reppermund, S., Thum, S., and Unger, K. (2010). Neuropsychological profiles in mci and in depression: Differential cognitive dysfunction patterns or similar final common pathway disorder? *Journal of psychiatric research*, 44(10):647–654.