What Difference Does it Make? Early Dementia Detection Using the Semantic and Phonemic Verbal Fluency Task

Hali Lindsay[1], **Johannes Tröger**[1], **Jan Alexandersson**[1], **Alexandra König**[2]
German Research Center for Artificial Intelligence (DFKI) Stuhlsatzenhausweg 3, Saarbrücken, Germany, 66125[1]
Memory Clinic, Association IA, CoBTek Lab CHU Université Côte d'Azur, France[2]
Hali.Lindsay@dfki.de

Abstract

Verbal Fluency (VF) tasks are common cognitive tests that are used in the diagnosis of early stages of Dementia. There are two main types of VF tasks; Semantic Verbal Fluency (SVF) and Phonemic Verbal Fluency (PVF). While much work has been done on automatic diagnostic relevance of the SVF, research on the automatic analysis of the PVF task or a combination of both remains minimal. This paper explores methods of extracting features from the SVF and the PVF task according to clinical and temporal methods, as well as how combined within-subject features from both tasks can increment classification performance. We investigate an early diagnostic scenario with a binary classification between healthy controls (N=8) and those with mild cognitive impairment (N=19), a likely precursor to dementia. Synthetic data augmentation (SMOTE) is used to balance the data set and multiple machine learning models— logistic regression, support vector machines with linear and radial basis function, and a multi-layer perceptron—are used to evaluate the features. The best performance comes from combining SVF, PVF and novel joint within-subject features (AUC > 0.90) for multiple machine learning methods.

Keywords: Phonemic Verbal Fluency, Semantic Verbal Fluency, Machine Classification, Clustering, mild cognitive impairment

1. Introduction

Verbal fluency (VF) tasks are amongst the most widely applied neuropsychological tests for the assessment of neurocognitive disorders. They are especially used for the diagnosis of different stages of dementia, ranging from very mild or even prodromal forms to clinical forms like in Alzheimer's disease. The main strength of VF tasks are their ease of use (no testing material required and fully speech-based interaction) and brevity (1-2 minutes) given a high sensitivity for above-mentioned diagnostic purposes. Despite their traditional wide adoption in clinical and diagnostic practice, there is an ongoing scientific discussion regarding what verbal fluency tasks actually measure in terms of neurocognitive functions. However, multiple studies show that VF tasks generate rich variance stemming from the interplay of multiple neurocognitive functions including executive functions (EF) as well as memory and language components. Differentiating between them and identifying those VF contributors is crucial to understand how VF could be used to differentiate between multiple dementia sub-forms. VF as a test category comprises two major versions, the semantic verbal fluency (SVF) and the phonemic verbal fluency (PVF). Both follow similar rules: One has to produce as much different words as possible within a given timeframe and a given constraint. In the SVF the constraint is that all produced words should belong to one semantic category (e.g. animals) and within the PVF the constraint is that all produced words should start with one letter (e.g. S).

Methodologically, it is best clinical practice to test for both VF: SVF and PVF. In this context, multiple studies report a Verbal Fluency Discrepancy, meaning a performance advantage in the SVF compared to the PVF. This often is explained by the fact that in SVF one can follow associations for word production, whereas In PVF associations additionally have to be monitored for their phonemic fit which puts additional EF demands on the testee. Therefore, in general, performance of healthy elderlies is better in the SVF than in PVF and this effect is preserved over aging (Vaughan et al., 2016). Patients with Mild Cognitive Impairment (MCI) of the amnestic type (precursor of Alzheimer's Disease) typically have less of a Verbal Fluency Discrepancy, meaning less of an advantage of SVF over PVF. Those patients that have a PVF advantage over the SVF are highly suspected to convert to AD (Vaughan et al., 2016; Teng et al., 2013). Conversely, a primary impairment in the PVF task is often regarded as strong indicator for EF impairment and frontal-lobe degeneration indicating Dementia of the fronto-temporal type (Dubois et al., 2000). When it comes to AD there seems to be an overall agreement that the SVF is more sensitive for conversion from different stages (e.g. healthy, MCI AD) than PVF (Alegret et al., 2018; Murphy et al., 2006; Amieva et al., 2005).

Automatic qualitative analysis of both the SVF and the PVF using computational linguistics methods has shown very strong results in modelling strategy usage in this tasks (Lindsay et al., 2019; Tröger et al., 2019; Linz et al., 2017b) and ultimately also in automatically classifying between multiple cognitive disorders (König et al., 2018). However, research rarely takes into account both tasks or automatically calculates features to model the Verbal Fluency Discrepancy well-reported in clinical research.

Therefore the aim of this paper is to take advantage of the complementary diagnostic power of both VF tasks, combine this with previously established qualitative computational linguistics methods and prove such an approach's overall quality for automatic classification in a traditionally very difficult applied machine learning scenario: Mild Cognitive Impairment vs. Healthy Controls.

2. Background

Traditionally, VF tasks are evaluated by counting all the relevant words produced in the given time frame, excluding repetitions. Although intrusions and repetitions have been investigated, state-of-the-art clinical evaluation of VF tasks centers around basic quantitative measures modelling neither qualitative aspects nor the temporal fine-grained resolution of a VF production.

2.1. Qualitative Evaluation of Verbal Fluency

Neuropsychological research investigated the quality of VF production early on by proposing a hierarchical set of rules to define qualitatively connected parts of a production (1997). The motivating rational being that people do not produce words randomly but rather produce spurts of related words, or clusters. When a person has run out of easily accessible words, they intentionally navigate to a new associative field and exploit words from there, generally referred to as switching.

These early efforts propose for the SVF task a taxonomic approach with pre-defined semantic sub-categories (e.g. SVF on animals with subcategory farm animals or African animals). For the PVF task, a rule-based system is used to determine phonemic associations by manually defining criteria for phonemic similarity (Vonberg et al., 2014; Troyer et al., 1997); e.g. for PVF on the letter A, words are scored as associated/connected if they share common first letters like arm & art.

More recently multiple computational approaches have been proposed to model similar qualitative aspects within a VF performance. Approaches include inferring semantic associations in the SVF through distributional semantics (Linz et al., 2017a; Woods et al., 2016; Pakhomov and Hemmy, 2014), language models (Linz et al., 2018) or graph theory (Clark et al., 2016) and with data-driven approaches for multiple variations of edit distances in PVF (Lindsay et al., 2019).

Although automatic approaches to model qualitative aspects of VFs have been shown to be promising for both classification scenarios as well as classic inferential statistics experiments, they remain experimental and appear to be chosen subjectively if not arbitrarily.

Within this study, we will use automatic implementations of the rule-based early methods proposed by Troyer to keep results between the tasks comparable, as only this rule-based framework models qualitative aspects of SVF and PVF alike.

2.2. Temporal Evaluation of Verbal Fluency

While qualitative aspects of VF productions have been studied early on, temporal fine-grained modelling of such a production has been studied only recently. One main reason might be that temporal analysis of VF can only be performed if patients' productions are recorded and transcribed. Tröger et al. (2019), suggests a temporal approach in which words that are said in close succession are considered to be in a cluster, regardless of semantic or phonemic motivation. The rationale behind it is that words which are—for whatever reason—associated in a person's semantic memory will be more accessible hence produced in faster succession.

While this has been used for the evaluation of the SVF task, there is—as of writing this paper—no research on the behavior of temporal clustering on the PVF task.

We choose to use this as one of our methods of feature generation as it does not require a semantically or phonemically motivated reason behind clustering and allows for an equal opportunity approach to both tasks.

2.3. Binning Approach to Verbal Fluency

Linz et al(2019) proposed a different temporal method for analyzing verbal fluency tasks where a one-minute speech sample is cut in to six 10-second bins. Features can then be calculated from each of the 10 second bins allowing for a finer resolution of the features over the task. In this paper, they looked at the word count, transition length and clustering features by bin. They found promising results using this technique in classifying between HC and MCI in Swedish subjects with the SVF Task, specifically for the word count of the last two intervals, 40-50 seconds and 50-60 seconds. These findings were supported by correlating different binning features with other trusted neuro-psychological tests such as the Boston Naming Test.

2.4. Automatic Classification of Verbal Fluency Tasks

Making use of novel qualitative as well as temporal features from VFs, recent work on automatic classification scenarios yields promising results. (Ryan, 2013) used logistic regression and a combination of SVF and PVF features to classify between healthy controls (HC) and MCI yielding an AUC of 0.76. (Linz et al., 2019) used temporal features extracted from SVF to classify between Swedish HC and MCI with the best result of an AUC of 0.72. Earlier in a classification experiment on French SVF data from HC and MCI (Linz et al., 2017a) achieved an F1 score of 0.77 by using qualitative semantic features.

	HC	**MCI**	***p***
N	8	19	
Sex (M/F)	8/0	12/7	-
Age (years)	71.50 (7.33)	75.32 (6.26)	0.18
Education (years)	9.75 (4.83)	10.53 (4.10)	0.67
MMSE (max 30)	29.25(0.89)	25.53 (3.31)	≤ 0.001

Table 1: Demographic information for the French population used in this analysis.

3. Methods

3.1. Data

The data used in this research was collected during the Dem@Care (Karakostas et al., 2017) and ELEMENT (Tröger et al., 2017) projects. Participants were recruited through the Memory Clinic located in Nice University Hospital at the Institute Claude Pompidou. Data was collected in the form of speech recordings via an automated recording application installed on a tablet computer. The recordings were manually transcribed in PRAAT (Boersma

Feature Name	Description
Word Count	The total number of animal words said in one minute, excluding repetitions
Mean Latency	Mean time (in seconds) elapsed since first utterance over all words
Troyer Measures	
Mean Troyer Cluster Size	Average number of animals in an SVF cluster over the entire sample
Number of Troyer Switches	the number of switches between Troyer clusters
Temporal Measures	
Mean Temporal Cluster length	Mean time (in seconds) spent inside a cluster.
Mean Temporal Cluster Coherence	Mean time (in seconds) spent between words inside clusters
Mean Temporal Cluster Size	the mean number of words inside a temporal cluster.
Number of Temporal Switches	the number of switches between temporal clusters
Mean Temporal Switch Coherence	Mean time (in seconds) between any two consecutive clusters.
Bin Measures	
Word Count by Bin	The number of words per 10 second bin
Transition Length by Bin	The average transition time in seconds between the end of one word and the onset of the next word by 10 second bin
Difference Measures	
Word Count	The total number of animal words said in one minute, excluding repetitions
Mean Latency	Mean time (in seconds) elapsed since first utterance over all words
Mean Troyer Cluster Size	Average number of animals in an SVF cluster over the entire sample
Number of Troyer Switches	the number of switches between Troyer clusters
Mean Temporal Cluster length	Mean time (in seconds) spent inside a cluster.
Mean Temporal Cluster Size	the mean number of words inside a temporal cluster.
Number of Temporal Switches	the number of switches between temporal clusters
Mean Temporal Switch Coherence	Mean time (in seconds) between any two consecutive clusters.
Word Count by Bin	Difference between SVF and PVF word count for respective 10s bin
Transition Length by Bin	SVF - PVF average transition in seconds between words for respective 10s bin

Table 2: The following features were extracted from the SVF and PVF task produced by the participants. For a more detailed explanation of how clusters are determined, please see Section 3.2.

.

and Weenink, 2009) according to the CHAT protocol (MacWhinney, 1991). Participants were asked to complete a battery of cognitive tests, including a 60 second semantic verbal fluency task—on the topic animals—and a 60 second phonemic verbal fluency task—for the letter category *F*. Demographics for the data used are displayed in 1 with significance testing between the populations. The MMSE (mini mental state examination) is a widely used test for cognitive performance where performance is measured on scale of 0 to 30 where anything below 25 is considered to be a sign of impairment. For this analysis four outliers were removed so that the maximum age considered was 85 years. One HC was removed for having an MMSE of 25, which would typically reflect some form of impairment.

3.2. Features

To investigate the diagnostic power of the SVF and PVF tasks, we designed three unique feature sets.

The first two are created by looking at each task individually; a *SVF feature set* and *PVF feature set*. An identical set of features were extracted from each task, according to how the task is evaluated. For example, the word count feature for the SVF is the number of animals said during the task, excluding repetitions and the for PVF it is the number of words starting with the letter *F* produced during the task, excluding repetitions. Word count, mean latency, Troyer measures, temporal measures and bin measures are extracted from each participant file for the single-task features sets and are described in the top-half of Table 2. A third feature set is created by combining the tasks by subtracting the PVF feature values from the corresponding SVF feature values of the same patient. This is referred to as the *difference features set*. For a detailed list of features and how they were produced see Table 2.

For the automatic computation of the Troyer clusters, SVF clusters are implemented according to the methodology in (Linz et al., 2017a) where a hierarchical set of predefined rules is used to determine semantically motivated clusters. Automatic Phonemic clusters, for the phonemic verbal fluency task are achieved by automating the phonetic rules proposed by Troyer as done in (Lindsay et al., 2019). Temporal clusters were computed according to Tröger et al. (2019). Binning measures for word count and transition length are calculated according to (Linz et al., 2019). As an additional temporal evaluation we computed mean latency, the average response latency for each word calculated by measuring the elapsed time since the onset of the first uttered word (Rohrer et al., 1995).

3.3. Oversampling with SMOTE

Due to a small and unbalanced data-set, the Synthetic Minority Oversampling Technique (SMOTE) is used to balance the HC and MCI population during the training phase of the classification experiments. This technique oversam-

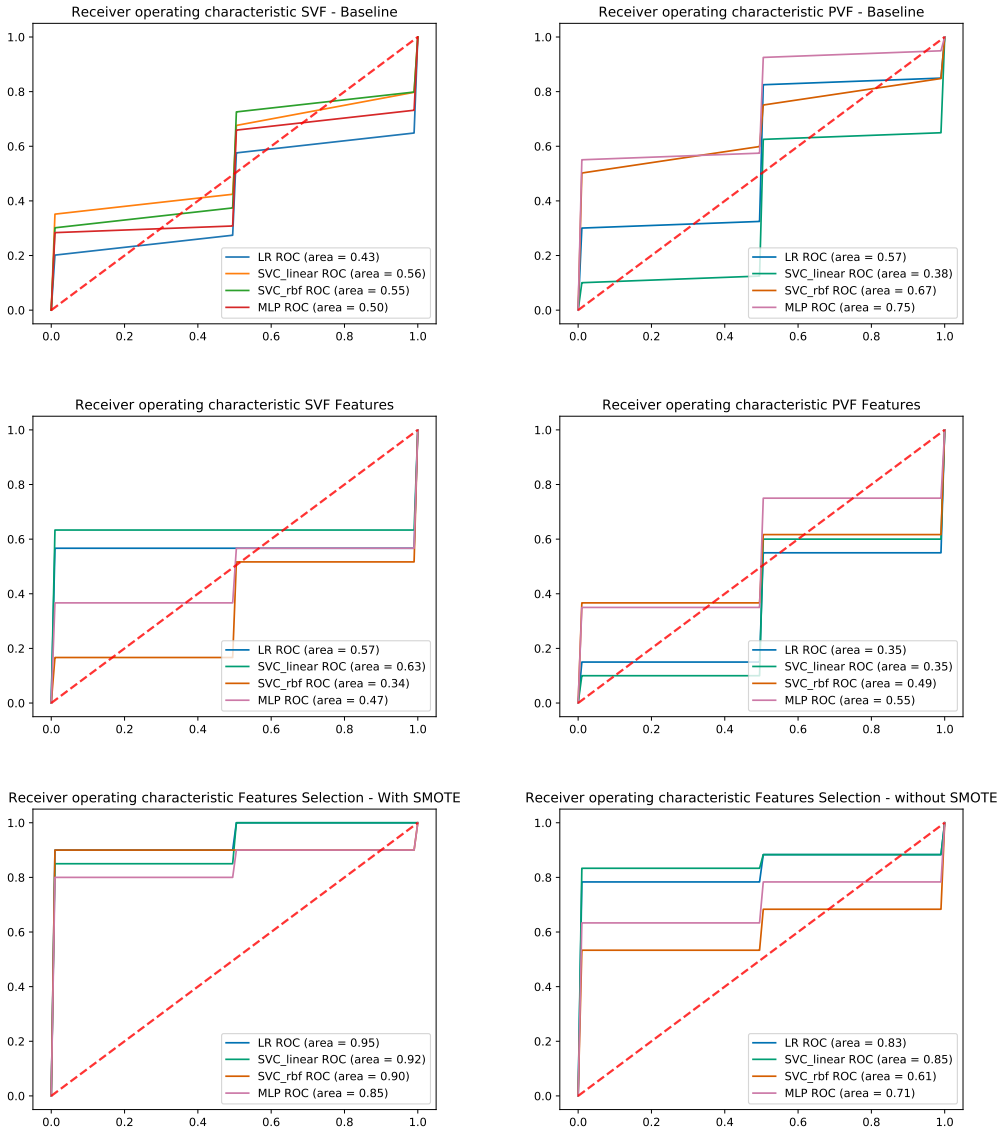

Figure 1: ROC Curves and corresponding AUC values for the HC vs. MCI classification experiments. Models are indicated by color. The red dashed line represents chance performance.

ples the minority class—in this case, healthy controls—by creating synthetic healthy samples so that the data becomes balanced. Synthetic samples are created by randomly choosing an existing example and drawing a line to other similarly existing samples within the features space. New synthetic samples are then randomly chosen points from this line (Bowyer et al., 2011). SMOTE is applied during the cross-fold validation, after splitting the fold data into training and testing sets, SMOTE is applied only on the training data. The testing split is not augmented. Therefore, testing is only done on real samples.

3.4. Classification Experiments

To consider the validity of this approach, a series of machine learning experiments are conducted with the different proposed features sets, simulating a screening scenario: automatically differentiating between healthy controls and patients with mild cognitive impairment. The focus of this paper is on the application of looking at the feasibility of combining PVF and SVF for early diagnosis of Dementia.

Six experimental scenarios were created from the features described in Section 3.2.: 1) an SVF Clinical Baseline where only the SVF word count is considered, 2) a PVF Clinical Baseline where only the PVF word count is considered, 3) the SVF feature set, 4) the PVF feature set, 5) the SVF, PVF and DIFF feature set and 6)the SVF, PVF and DIFF feature set without SMOTE. To obtain a more comprehensive picture, four machine learning approaches are considered:

- A Logistic Regression (LR) model is created where error is minimized by least square errors, what is commonly referred to as the L2 loss.
- A Multilayer Perceptron (MLP) model is used with the Limited-memory BFGS (lbfgs) solver and logistic activation function. The alpha parameter is set at 0.1 and an adaptive learning rate is used.

- A Support Vector Machine Classification model is also created with a linear (SVC-linear) kernel.
- A Support Vector machine Classification with a radial basis function (SVC-rbf) kernel.

For each experiment, 5-fold cross validation is used All models are created simultaneously so that the same split of the data in each fold is used to train and test each of the four models. No parameter optimization is used. For scenarios 5) the SVF, PVF and Difference feature set with SMOTE and 6)the SVF, PVF and DIFF feature set without SMOTE, features selection is used in the training phases where and independent t-test is used to determine the significance of the features between the groups. Features with a p-value greater than 0.05 are discarded and the remaining features are used to train. The classification models are created using the scikit-learn library in Python3 (Pedregosa et al., 2011).

4. Results

4.1. Classification Results

Results from the classification experiment are displayed in Table 3. For evaluation, accuracy, sensitivity, precision , F1 score and Area Under the Receiver Operator Curve (AUC) are provided. The mean score from the 5-fold cross validation is given as well as the standard deviation in parentheses. Each feature set is displayed with the results from each model described in Section 3.4.. Results are visualized with receiver operator characteristic (ROC) curves in Figure 1, where a larger area under the curve (AUC) indicates that the model is better at differentiating between HC and MCI.

During the feature selection process, the features that were chosen based on their significance were *PVF word* $count_{40-50}$, *SVF word* $count_{40-50}$, *SVF transition* $length_{40-50}$, *DIFF word* $count_{10-20}$.

From the clinical baseline models, where just word count is used, SVF showed consistent AUCs of roughly 0.5 across models, with the highest AUC of 0.56 coming from SVC-Linear. For PVF, AUC scores varied in the baseline from a low of 0.38 by the SVC-Linear and the highest AUC of 0.75 coming from the MLP. Using all the automated SVF features, SVF improved from the baseline, achieving its highest score with SVC-Linear at 0.63. An increase of 0.07 from the SVF baseline. PVF decreased with the additional automated features from the word count baseline with its highest AUC reaching 0.55. The best results are found when combining the SVF, PVF and DIFF features. The highest AUC of 0.95 is achieved by LR. The lowest AUC of 0.85 is achieved by the MLP. The Accuracy of these models is maintained between 70 and 80%.

We consider this scenario without SMOTE to see how oversampling might affect the training. There is a slight dip in performance. However it exceeds all other SVF and PVF models. The highest AUC of 0.86 is achieved by SVC-Linear at 0.86. The lowest AUC is found at 0.61 by the SVC-RBF. However, the accuracy is, on average, higher than with SMOTE. The average accuracy across models with SMOTE is 74.8% but without SMOTE it is 75.5%.

5. Discussion

Looking at the results from the classification models, there is improvement from the clinical baseline of word count from both the SVF and PVF task in comparison to the all features model both with and without SMOTE, which hinges on the additional computational measures.
The classification for both the SVF-baseline (AUC=0.56) and all SVF features (AUC=0.63) seems low. Previous papers reported AUCS of over 0.7 using similar features and models to distinguish between HC and MCI (Linz et al., 2019)(Linz et al., 2017b). One difference that may have lead to this result is using a hierarchical predefined list to determine clusters instead of using an automated approach, such as clustering using semantic word embeddings. We also chalk this low result up to the relative size of the data set.

However, for SVF there is at least some improvement over the baseline using the additional computational measures. This is not mirrored by the PVF task where the baseline (AUC = 0.75) is better than the additional features (AUC = 0.55). PVF lacks the foundation of research in computational measures. From a feature standpoint, previous methods that have shown to be beneficial for evaluating the SVF task seem to transfer to the PVF task (e.g. clustering, binning, ect). Future work should look at the underlying production strategies and cognitive processes engaged, during the PVF, similar to the work that has been done on SVF, in order to improve its classification as a standalone task.
An interesting finding from the feature selection is that features from each data set were used to achieve the best classification result; PVF word $count_{40-50}$, SVF word $count_{40-50}$, SVF transition $length_{40-50}$, DIFF word $count_{10-20}$. This highlights the need for diverse features even in a small data setting. (Linz et al., 2019) also found similar results at the 40-50s bin for SVF in an early diagnosis scenario for dementia. They found that SVF word count in the 40-50s bin correlated positively with scores for other neuro-psychological tests that measure vocabulary, namely the Boston Naming Test and WAIS similarity task. It is interesting to see that this finding is repeated in the PVF task. This is something that should be investigated for the PVF task with a larger data set. It should also be considered for further investigations of underlying cognitive processes. This result also highlights the benefits of using binning as a qualitative temporal analysis method of the verbal fluency tasks.

While some of the machine learning classification results presented are quite accurate, this may be due to the synthetic data augmentation technique, SMOTE. While clinical data tends to be relatively noisy, the synthetic data is probably a cleaner data set than real life conditions. To test this, we need more data and a balanced data set to confirm that the classification results presented here hold in real-world testing conditions. While we do expect slightly worse results without SMOTE—as shown in

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
Semantic Verbal Fluency - Clinical Baseline						
LR	0.48(0.11)	0.53(0.17)	0.30(0.24)	0.65(0.08)	0.58(0.14)	0.42(0.26)
SVC - Linear	0.53(0.17)	0.62(0.17)	0.30(0.24)	0.68(0.10)	0.64(0.14)	**0.56(0.27)**
SVC - RBF	0.43(0.14)	0.47(0.17)	0.30(0.24)	0.62(0.10)	0.53(0.15)	0.55(0.23)
MLP	0.56(0.13)	0.65(0.25)	0.40(0.37)	0.75(0.14)	0.65(0.16)	0.50(0.10)
Phonemic Verbal Fluency - Clinical Baseline						
LR	0.49(0.24)	0.45(0.40)	0.70(0.24)	0.44(0.36)	0.44(0.37)	0.57(0.29)
SVC - Linear	0.68(0.19)	0.77(0.29)	0.50(0.32)	0.81(0.11)	0.75(0.17)	0.38(0.27)
SVC - RBF	0.86(0.07)	0.95(0.10)	0.60(0.37)	0.88(0.10)	0.90(0.05)	0.68(0.26)
MLP	0.79(0.14)	0.85(0.12)	0.60(0.37)	0.86(0.12)	0.85(0.09)	**0.75(0.21)**
Semantic Verbal Fluency Features						
LR	0.55(0.15)	0.58(0.25)	0.40(0.37)	0.76(0.13)	0.62(0.16)	0.57(0.28)
SVC - Linear	0.55(0.10)	0.58(0.25)	0.40(0.49)	0.80(0.16)	0.62(0.13)	**0.63(0.25)**
SVC - RBF	0.51(0.17)	0.58(0.19)	0.30(0.24)	0.67(0.09)	0.61(0.16)	0.34(0.24)
MLP	0.48(0.11)	0.63(0.19)	0.10(0.20)	0.63(0.07)	0.62(0.10)	0.47(0.32)
Phonemic Verbal Fluency Features						
LR	0.56(0.13)	0.65(0.30)	0.40(0.37)	0.75(0.13)	0.64(0.17)	0.35(0.22)
SVCl- Linear	0.55(0.10)	0.63(0.25)	0.40(0.37)	0.75(0.13)	0.64(0.14)	0.35(0.18)
SVC - RBF	0.66(0.18)	0.70(0.29)	0.60(0.37)	0.84(0.13)	0.71(0.20)	0.49(0.31)
MLP	0.56(0.13)	0.65(0.30)	0.40(0.37)	0.75(0.13)	0.64(0.17)	**0.55(0.33)**
All Features with Feature Selection - With SMOTE						
LR	0.77(0.14)	0.78(0.19)	0.80(0.24)	0.91(0.11)	0.82(0.11)	**0.95(0.10)**
SVC - Linear	0.71(0.12)	0.73(0.16)	0.70(0.24)	0.84(0.13)	0.77(0.11)	0.93(0.10)
SVC - RBF	0.79(0.18)	0.85(0.20)	0.70(0.40)	0.88(0.15)	0.84(0.13)	0.90(0.12)
MLP	0.72(0.18)	0.80(0.19)	0.50(0.32)	0.79(0.11)	0.79(0.14)	0.85(0.12)
All Features with Feature Selection - without SMOTE						
LR	0.78(0.06)	0.80(0.10)	0.70(0.40)	0.90(0.12)	0.84(0.04)	0.83(0.14)
SVC - Linear	0.76(0.13)	0.83(0.14)	0.60(0.37)	0.85(0.13)	0.83(0.08)	**0.86(0.13)**
SVC - RBF	0.74(0.07)	0.95(0.10)	0.20(0.24)	0.75(0.05)	0.84(0.05)	0.61(0.24)
MLP	0.74(0.09)	0.90(0.12)	0.30(0.40)	0.79(0.11)	0.83(0.05)	0.71(0.12)

Table 3: This table contains results from the classification experiments. All Features is the combination of SVF, PVF and difference features. For evaluation, accuracy, sensitivity, precision , F1 score and Area Under the Receiver Operator Curve (AUC) are provided. The mean of the 10-fold validation is given. The standard deviation is given in parentheses. Highest Accuracy and AUC scores are emphasized in bold font. All models use SMOTE except for the final experiment which is labeled *All Features - Without SMOTE*.

the All Features without SMOTE classification experiment (SVC-Linear AUC=0.86)—we still expect improved results with more authentic data. It is worth noting that SMOTE is not being used to increase the accuracy or AUC of the model directly, but is being used to improve the training data, which could indirectly influence the models' behavior.

6. Conclusion

In this paper, we explored using previously automated qualitative analysis techniques from the SVF—semantic and temporal clustering as well as temporal binning—on the PVF tasks with promising results. Moreover, as clinical research suggests, we present an approach to fuse both tasks in calculating specific difference features harnessing the so-called Verbal Fluency Discrepancy in AD and its precursor stage MCI. The features generated from both of these tasks as well as the development of multi-task joint difference features lead to improved classification for early detection of Dementia symptoms and are verified by multiple classifiers, with and without synthetic data augmentation to balance a small clinical data set. While the results are promising, this paper setup a pipeline for the feasibility of creating classification experiments with multiple verbal fluency tasks. This work should be reiterated in other languages as well as on larger data sets to confirm the suggested conclusions.

7. Bibliographical References

Alegret, M., Pereto, M., Perez, A., Valero, S., Espinosa, A., Ortega, G., Hernandez, I., Mauleón, A., Rosende-Roca, M., Vargas, L., et al. (2018). The role of verb fluency in the detection of early cognitive impairment in alzheimer's disease. *Journal of Alzheimer's Disease*, 62(2):611–619.

Amieva, H., Jacqmin-Gadda, H., Orgogozo, J.-M., Le Carret, N., Helmer, C., Letenneur, L., Barberger-Gateau, P., Fabrigoule, C., and Dartigues, J.-F. (2005). The 9 year cognitive decline before dementia of the alzheimer

type: a prospective population-based study. *Brain*, 128(5):1093–1101.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1.13).

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority oversampling technique. *CoRR*, abs/1106.1813.

Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R. M., Kapur, P., DeRamus, T. P., and Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement (Amst)*, 2:113–122.

Dubois, B., Slachevsky, A., Litvan, I., and Pillon, B. (2000). The fab. *Neurology*, 55(11):1621–1626.

Karakostas, A., Briassouli, A., Avgerinakis, K., Kompatsiaris, I., and Tsolaki, M. (2017). The dem@care experiments and datasets: a technical report. *CoRR*, abs/1701.01142.

König, A., Linz, N., Töger, J., Wolters, M., Alexandersson, J., and Robert, P. (2018). Fully automatic analysis of semantic verbal fluency performance for the assessment of cognitive decline. *Dementia and Geriatric Cognitive Disorders*. Accepted.

Lindsay, H., Linz, N., Tröger, J., and Alexandersson, J. (2019). Automatic data-driven approaches for evaluating the phonemic verbal fluency task with healthy adults. In *ICNLSP*.

Linz, N., Tröger, J., Alexandersson, J., and König, A. (2017a). Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Linz, N., Tröger, J., Alexandersson, J., Wolters, M., König, A., and Robert, P. (2017b). Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728, Nov.

Linz, N., Tröger, J., Lindsay, H., König, A., Robert, P., Peter, J., and Alexandersson, J. (2018). Language modelling for the clinical semantic verbal fluency task. In Dimitrios Kokkinakis, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., and Kokkinakis, D. (2019). Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 103–113, Minneapolis, Minnesota, June. Association for Computational Linguistics.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Inc.

Murphy, K. J., Rich, J. B., and Troyer, A. K. (2006). Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of alzheimer's type dementia. *Journal of the International Neuropsychological Society*, 12(4):570–574.

Pakhomov, S. V. and Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106.

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.

Rohrer, D., Wixted, J. T., Salmon, D. P., and Butters, N. (1995). Retrieval from Semantic Memory and its Implications for Alzheimer's Disease. *J Exp Psychol Learn Mem Cogn*, 21(5):1127–1139.

Ryan, J. (2013). *A System for Computerized Analysis of Verbal Fluency Tests*. Ph.D. thesis, 06.

Teng, E., Leone-Friedman, J., Lee, G. J., Woo, S., Apostolova, L. G., Harrell, S., Ringman, J. M., and Lu, P. H. (2013). Similar Verbal Fluency Patterns in Amnestic Mild Cognitive Impairment and Alzheimer's Disease. *Archives of Clinical Neuropsychology*, 28(5):400–410, 06.

Tröger, J., Linz, N., Alexandersson, J., König, A., and Robert, P. (2017). Automated Speech-based Screening for Alzheimer's Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

Troyer, A. K., Moscovitch, M., and Winocur, G. (1997). Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults. *Neuropsychology*, 11(1):138–146.

Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., and Kray, J. (2019). Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia*, 131, 05.

Vaughan, R. M., Coen, R. F., Kenny, R., and Lawlor, B. A. (2016). Preservation of the semantic verbal fluency advantage in a large population-based sample: Normative data from the tilda study. *Journal of the International Neuropsychological Society*, 22(5):570–576.

Vonberg, I., Ehlen, F., Fromm, O., and Klostermann, F. (2014). The absoluteness of semantic processing: Lessons from the analysis of temporal clusters in phonemic verbal fluency. 9:e115846, 12.

Woods, D. L., Wyma, J. M., Herron, T. J., and Yund, E. W. (2016). Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, 11(12):1–37.