Generating Synthetic Clinical Speech Data Through Simulated ASR Deletion Error

Hali Lindsay[1], Johannes Tröger[2], Mario Mina[2], Nicklas Linz[2], Philipp Müller[1], Jan Alexandersson[1], Inez Ramakers[3]

[1]German Research Center for Artificial Intelligence (DFKI) Stuhlsatzenhausweg 3, Saarbrücken, Germany, 66125

[2]ki:elements, Am Holzbrunnen, Saarbrucken,Germany, 66121

[3]Maastricht University Medical Center (MUMC+)

{Hali.Lindsay, Philipp.Mueller, Jan.Alexandersson}@dfki.de, {Johannes.Troeger, Mario.Mina, Nicklas.Linz}@ki-elements.de, i.ramakers@maastrichtuniversity.nl

Abstract

Training classification models on clinical speech is a time-saving and effective solution for many healthcare challenges, such as screening for Alzheimer's Disease over the phone. One of the primary limiting factors of the success of artificial intelligence (AI) solutions is the amount of relevant data available. Clinical data is expensive to collect, not sufficient for large-scale machine learning or neural methods, and often not shareable between institutions due to data protection laws. With the increasing demand for AI in health systems, generating synthetic clinical data that maintains the nuance of underlying patient pathology is the next pressing task. Previous work has shown that automated evaluation of clinical speech tasks via automatic speech recognition (ASR) is comparable to manually annotated results in diagnostic scenarios even though ASR systems produce errors during the transcription process. In this work, we propose to generate synthetic clinical data by simulating ASR deletion errors on the transcript to produce additional data. We compare the synthetic data to the real data with traditional machine learning methods to test the feasibility of the proposed method. Using a dataset of 50 cognitively impaired and 50 control Dutch speakers, ten additional data points are synthetically generated for each subject, increasing the training size for 100 to 1000 training points. We find consistent and comparable performance of models trained on only synthetic data (AUC=0.77) to real data (AUC=0.77) in a variety of traditional machine learning scenarios. Additionally, linear models are not able to distinguish between real and synthetic data.

Keywords: Data Augmentation, Synthetic Data, Clinical Speech, Mild Cognitive Impairment, Automatic Speech Recognition, Machine Learning

1. Introduction

Analysing clinical speech by means of natural language processing (NLP) techniques is a low-cost and effective approach for many healthcare challenges, such as screening for early signs of Alzheimer's Disease from clinical speech tasks. One of the primary limiting factors of the success of artificial intelligence (AI) solutions in health is the amount of relevant data available to train models. Clinical speech data is expensive and invasive to collect and the quantity is not sufficient for large-scale machine learning or even simple neural methods. In addition, collected data is difficult—if not impossible—to share between clinical and research institutions due to concerns for patient privacy. With the increasing demand for digital AI-driven solutions in health systems, generating synthetic clinical data that can be scaled-up and performs on-par with real data is the next challenge.

Previous work has shown that automated evaluation of clinical speech tasks via automatic speech recognition (ASR) is comparable to manually annotated results in diagnostic scenarios even though ASR systems produce errors during the transcription process, namely deletion (König et al., 2018; Konig et al., 2019). While the ASR-related loss of data in such a setting is typically seen as one of the major limitations of those approaches, this natural limitation can be harnessed to naturally generate synthetic data. The concept is similar to a technique used for synthetic data augmentation in computer vision, where random parts of an image are erased in order to generate multiple training examples from a single image (Zhong et al., 2017). We propose a novel technique for synthetic data augmentation by exploiting the already occurring ASR error to randomly delete portions of the transcribed clinical speech.

In this paper, we investigate if the technique of randomly erasing speech transcripts—a result which is already seen when using ASR systems as part of an automatic pipeline—can be applied to clinical speech to generate synthetic training data. This is done using 100 older Dutch speakers where 50 show signs of mild cognitive impairment. Ten synthetic files are generated per participant for a total of 1000 data points. This paper is scoped to consider if the synthetically generated data has comparable results to authentic data in traditional machine learning scenarios. Through a series of downstream machine learning classification experiments, the synthetic data is compared to the traditional scenario as a baseline. Overall, we find that random erasing can be used to generate synthetic clinical data that performs as well as the real data. Based on the foundation of these findings, future work should investigate if more complex neural methods benefit from the addition of synthetic data as well as if the proposed method is transferable to other clinical tasks.

2. Background

In this section, background is provided to further motivate the argumentation of the paper. First, the automatic pipeline for evaluating clinical speech is described. Next, focusing on going from speech to text portion of the automatic evaluation pipeline, an explanation of how the quality of the transcription is estimated if provided . Finally, drawing from data augmentation techniques in computer vision, parallels are drawn between the technique of random erasing and the role of deletion during the transcription process.

2.1. Automatic Evaluation of the Semantic Verbal Fluency task (SVF)

The semantic verbal fluency task is a timed clinical speech test where a person is asked to name as many words as they can pertaining to a given semantic category (e.g. Name as many animals as you can in one minute). This task has been shown to be sensitive for screening for mild cognitive impairment (MCI) from typical ageing in older adults (McDonnell et al., 2020; Clark et al., 2009; Vaughan et al., 2016). The automatic pipeline for evaluating this speech task starts with recording a person during the task. Next, this speech is passed through an automatic speech recognition (ASR) model to obtain a text transcript. Once this automatic transcript has been generated multiple methods of feature extraction and analysis have been proposed for evaluating the SVF task based on relevant cognitive clinical literature. Previous work has investigated using semantically motivated measures, such as semantic word embeddings, to consider semantic clustering strategies (Troyer et al., 1997; Pakhomov and Hemmy, 2014). Other methods have considered temporal measures for clustering (Tröger et al., 2019) or investigating the task on a finer time resolution (Linz et al., 2019a).

2.2. Word Error rate (WER)

One of the first elements of an automatic pipeline for evaluating the SVF, is to automatically transcribe the speech task using automatic speech recognition. As with any automatic method, there is always some form of error. To evaluate automatic speech recognition, word error rate is used. Word error rate (WER) is the number of insertions, substitutions, and deletions that occur during the automatic transcription process divided by the number of words in the manual transcript. (Errattahi et al., 2018)

$$WER = \frac{Substitutions + Deletions + Insertions}{WordCount_{manualtranscript}}$$

The most common form of error found when automatically transcribing the SVF task is deletion. In addition, before extracting clinically relevant features from the task, the text is preprocessed, removing words outside the task domain. Therefore substitutions and insertions that are not in the semantic category (e.g. animals) would also be seen as deletions.

The effect of the automatic speech pipeline on this clinical task has been investigated previously by comparing manual versus automatic evaluation methods. Konig and colleagues found that both methods yielded comparable results when screening for dementia over the phone using the SVF (König et al., 2018).

2.3. Generating Synthetic Data

Drawing from computer vision, one of the common methods is to alter images in the training set by cropping, flipping, rotating, or randomly erasing part of the image. By perturbing the original image in some way, many versions of a single image can be created. Random erasing is a data augmentation technique where additional training data is created by erasing a random portion of an image in varying amounts. Although the idea is simple, it was previously proposed to reduce overfitting in deep learning image recognition models (Shorten and Khoshgoftaar, 2019; Zhong et al., 2017). This idea lends itself easily to the clinical speech application when combined with the WER caused by the automatic transcription process. Since the deletion caused by the WER does not affect the downstream application of detecting cognitive impairment from the speech recording, it should be possible to randomly delete portions of manual transcripts at the same rate as the WER. This can be done in many variations and combinations, yielding synthetically augmented data.

3. Data

100 older Dutch speakers completed a battery of cognitive tests including a one minute semantic verbal fluency on the subject of animals with a clinician from Maastricht University Clinic, Netherlands. Of the 100 participants, 50 are healthy controls (HC) and 50 present with mild cognitive impairment (MCI). The demographic data for the sample population is given in table 1.

	HC	**MCI**
N	50	50
Sex (M/F)	18/32	19/31
Age (years)	70.66 (8.96)	65.94 (7.80)
Word Error Rate (%)	20.29	23.13
MMSE (max 30)	28.68(1.27)	26.92 (2.07)

Table 1: Demographic information for the Dutch participants. The Mini-Mental State Exam (MMSE) is a test to measure cognitive function (Max score 30). Means are given with standard deviation in parentheses.

To complete the SVF task, participants are instructed to name as many animals as they can in one minute. The response is recorded and transcribed twice; once manually by trained clinicians via an iPad application. The second time the data is transcribed automatically

via Google translation services. In both cases, the responses are automatically pre-processed to remove any additional sounds, such as 'uhh' or 'ahh'. The final response result is a time-aligned list of animals. For example, a transcript could look like "dog, cat, lions, tiger, bear, blue whale, dolphin".

4. Methods

4.1. Data Augmentation by Random Erasing with WER (REWER)

First, the word error rate is calculated for each participant between the manually annotated and automatically generated transcript. The, number of words to be deleted from the transcript is determined given the WER percentage. Because the goal is to simulate the naturally occurring error in the ASR transcript, the WER produced by the ASR is applied to the manual SVF transcript. An exhaustive list is created of every possible variation of the SVF with the determined number of missing words from the manual transcript. From this list, a random number generator is used to randomly select ten newly generated, synthetically augmented SVF texts per manual transcript. These files are then saved for the next step of explicit feature extraction.

4.2. Data Augmentation by Random Erasing with Constant Deletion (REWCD)

One of the limiting factors of the REWER method is that it requires manual transcription of the SVF task in order to calculate the WER. To investigate additional random erasing methods that do not require manual annotation, a constant rate of deletion is considered. Instead of a variable per participant deletion rate based on the WER determined by ASR, a constant deletion rate is considered for all transcripts. The same procedure is applied as describe in Section 4.1 where the rate of deletion is 10% and 20%. These rates are chosen to be below the average WER for the sample population given in Table 1.

4.3. Feature Extraction

A comprehensive feature set is extracted from the automatic and augmented transcripts based on recent approaches for automatically evaluating the SVF task. Previous literature has proposed investigating the underlying strategy for completing the SVF task by looking for clusters of semantically related words in the task (Troyer et al., 1997; Farzanfar et al., 2018). This process was previously automated and Four features are extracted for semantic clustering and switching based on Linz et al., 2017 using pre-trained Dutch semantic word embeddings from Fasttext (Bojanowski et al., 2016). Beyond semantics, temporal methods have also been proposed for extracting five clustering and switching metrics in SVF based on (Tröger et al., 2018). In addition, another temporal method has been investigated by breaking the sixty second task into six ten-second bins (Linz et al., 2019b; Lindsay et al., 2021). For more detailed feature explanations, short descriptions of each feature are given in Table 2

5. Experiments

Statistical Analysis is done in R Studio (R Core Team, 2017). All coding experiments are implemented using python 3.7.

5.1. Machine learning Classification Scenarios

To test the feasibility of the proposed data augmentation technique, multiple machine learning experiments are conducted.

5.1.1. Augmentation Approach with REWER and REWCD

This train-test setup is applied to the three synthetic data sets that are created as well as the combination of all of them: the WER set (WER), the 10% constant rate (C_10%), the 20% constant rate (C_20%), and the combination of all three synthetic datasets (ALL SYNTH). Since the idea of this paper is to produce synthetic data that performs similarly to the real data, we propose to train on the synthetically generated data and test on the real ASR data. To keep the models comparable to training and testing on the real ASR data, leave one out cross validation(LOOCV) is also used in this scenario. In this case, the one participant that is in the test set has all their synthetic data removed from the training set. For a concrete example, this means of the 1000 generated files, 990 synthetic data points are in the training set and 1 real data point is in the test set. The 10 data points that are removed are the files generated by the one being tested. This is done to prevent inflating model results.

5.1.2. Classic Approach

To compare the newly proposed technique to traditional methods, model performance is considered when training on the real ASR data (REAL ASR) using LOOCV.

5.1.3. Machine learning Classification Specifications

The classification models are created using the scikit-learn library[1] (Pedregosa et al., 2011).

For the feasibility of this method, binary classification is done to distinguish between the MCI and control group. Three classification algorithms are considered; Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVC). Features are normalized using the standard scalar. Grid search is used to optimize model parameters in the training fold. In addition, univariate feature selection is done to test how increasing the number of features increases as the data increases. To gauge and compare model performance, accuracy and area under the receiver operator curve (AUC) are calculated.

[1] sklearn version==0.24.0 for python 3.7

Feature Name	Description
Word Count	The total number of animal words said in one minute, excluding repetitions
Semantic Clustering Measures	
Mean Cluster Size	Average number of animals in an semantic cluster over the entire sample
Number of Switches	the number of times switched to a different semantic cluster
Mean Intercluster Similarity	On average, how semantically related are the semantic clusters
Temporal Clustering Measures	
Mean Transition Time	Mean time (in seconds) between consecutive words
Mean Cluster Size	Average number of animals in a temporal cluster over the entire sample
Number of Switches	the number of times switched to a different temporal cluster
Mean Intercluster Similarity	On average, how semantically related are the temporal 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
Semantic Similarity by Bin	On average, how semantically related the words are by 10 second bin

Table 2: Features extracted from the SVF task produced by the participants with description.

5.2. Additional Experiments

A few other experiments are considered to examine the synthetically generated data. A random baseline is generated using the permutation test to see if the synthetic data can be distinguished from the real ASR data. In addition, incremental experiments are considered to see how the amount of synthetic data used in training effects the binary diagnostic classification experiment.

5.2.1. Permutation Test

To test if we can tell the difference between the synthetic and authentic data, permutation test is computed. A permutation test consists of obtaining a randomised baseline by training a linear model a series of times while permuting the target labels in question each time, removing any dependence between the input features and the mentioned target label. In this case, the target label is authentic or synthetic. The p-value represents the probability of obtaining the model accuracies we observe, assuming the that the null hypothesis is true. For this experiment, the null hypothesis is that there is no difference between the synthetic and authentic data. To test this, authentic and synthetic labels are randomly assigned to the transcripts. A linear model is trained and tested with the randomly assigned labels. Accuracy is used to determine model performance. This is permuted 1000 times for comparison. An empirical p-value is calculated by computing how many of the random models have a higher accuracy than the model trained on the true labels. The empirical p-value is calculated by taking the number of times performance falls within the random model score distribution divided by the total number of permutations. The p-value, in this case, represents how many of the random models have superior or comparable performance to the one trained on the actual experimental scenario. We report the p-value with statistical significance set to 0.05.

5.2.2. Incremental Experiments

In addition, the amount of synthetic data used to train a model is tested where the training amount is increased incrementally. A model where one synthetic data point per participant is trained, then a model where two synthetic data points per participants is trained and so on, until all ten points per participant are used. In this scenario, the machine learning scenario is simplified. A simple logistic regression using all extracted features is created with no hyperparameter optimization. As stated previously, LOOCV is used where the synthetic data is used to train and real ASR data is used to test and no data from the test participant is seen during training.

6. Results

6.1. Machine Learning Results

Model	N	Accuracy	AUC	Method
BEST ACC				
LR	13	0.74	0.76	REWER
SVC	22	0.75	0.76	REWER
RF	11	0.73	0.76	ASR
BEST AUC				
LR	13	0.69	0.77	ASR
SVC	22	0.69	0.77	ALL SYNTH
RF	11	0.73	0.76	ASR

Table 3: Best result for feasibility experiments for each classifier. N is the number of features. Method is which training data had the best score. The upper table is based on highest accuracy. The lower tables is based on highest AUC.

Results for the machine learning experiments explained in Section 5.1 are visualized in Figure 1. In addition

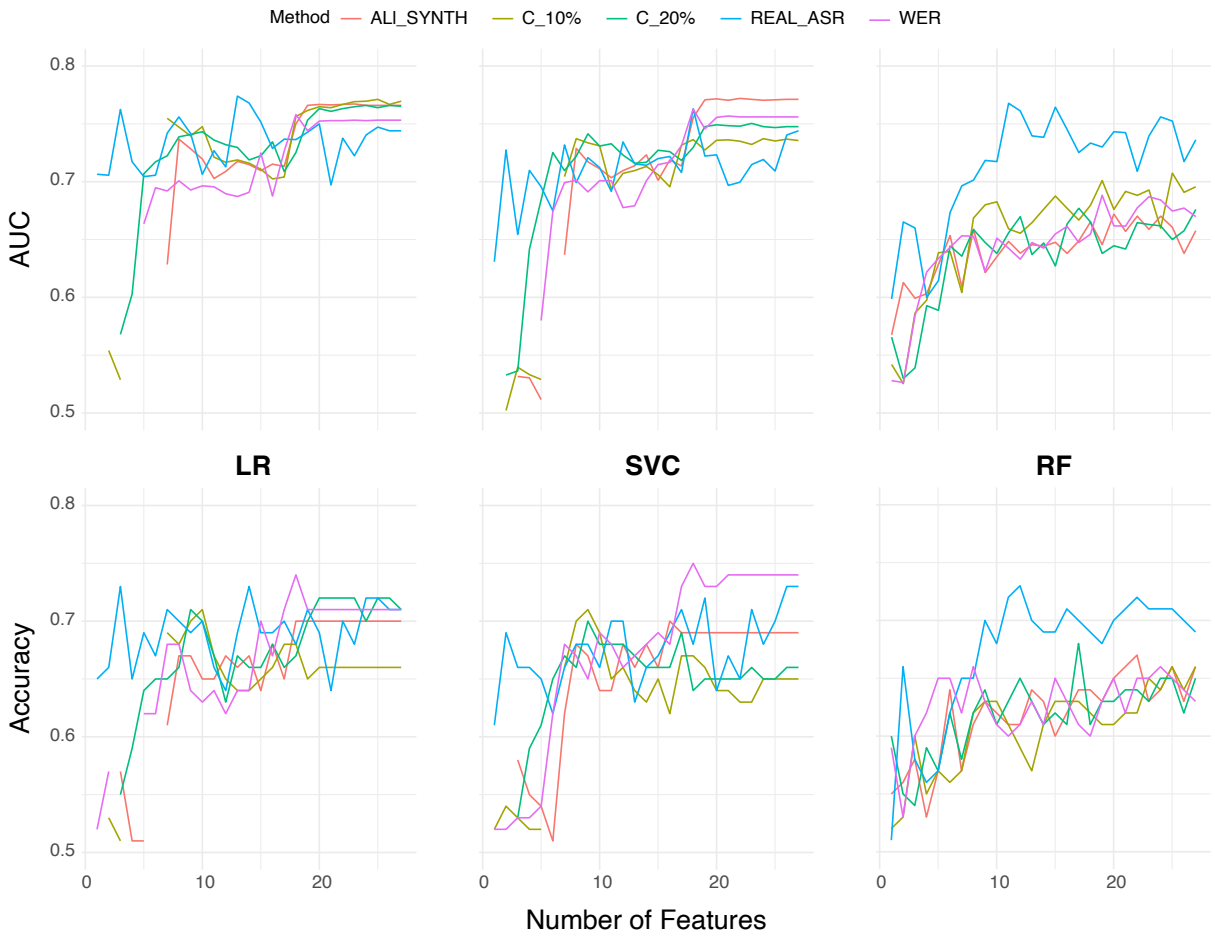

Figure 1: Visualization of results from the machine learning experiments. The number of features used to traain the modedl is represented on x-axis. Logistic Regression(LR). Support Vector Machine(SVC), Random Forest(RF). Area Under Curve (AUC).

the best accuracy and AUC score for each algorithm are displayed in Table 3.

From the results, comparable performance is seen between both the REWER and REWCD methods to the classic approach. Overall, the synthetic data improves in performance with the number of features. Looking at the best accuracy by classifier, for the logistic regression and support vector machine, the WER method produces the max result. For AUC, the support vector machine best result is achieved using the combined synthetic data sets. However, real data yields better AUC performance in general. There is also appears to be some dependence on classifier type the random forest classifier consistently performs better in both accuracy and AUC with real data.

6.2. Permutation Test Results

For the random baseline from the permutation tests, No significant p-values are reported. Values range from 0.44 to 0.56 with the average significance value being 0.51. Therefore, the alternate hypothesis is rejected and the null hypothesis is accepted. This can be interpreted as the linear model not being able to distinguish between the synthetic and authentic transcripts.

6.3. Incremental Experiments

Results for the incremental experiments are visualized in Figure 2. In addition, Table 4 summarizes the results by averaging AUC and accuracy scores by the number of synthetic data points used during training per participant.

As the amount of data increases, consistent AUC values are reported ranging from 0.74 to 077, and consistently averaging to 0.76. The accuracy presents with a mild downward slope 71% with one data to 69% accuracy at nine synthetic points per person. The slight decrease in accuracy could be to the lack of optimization during training as higher accuracy (74%) is reported for the logistic regression with ten data points per person.

7. Discussion

Of the synthetic methods considered, WER had the best accuracy. This result is expected based on the train-test setup used. The REWER method generates training data closest to the ASR test data. However, the constant deletion had comparable results to the WER and

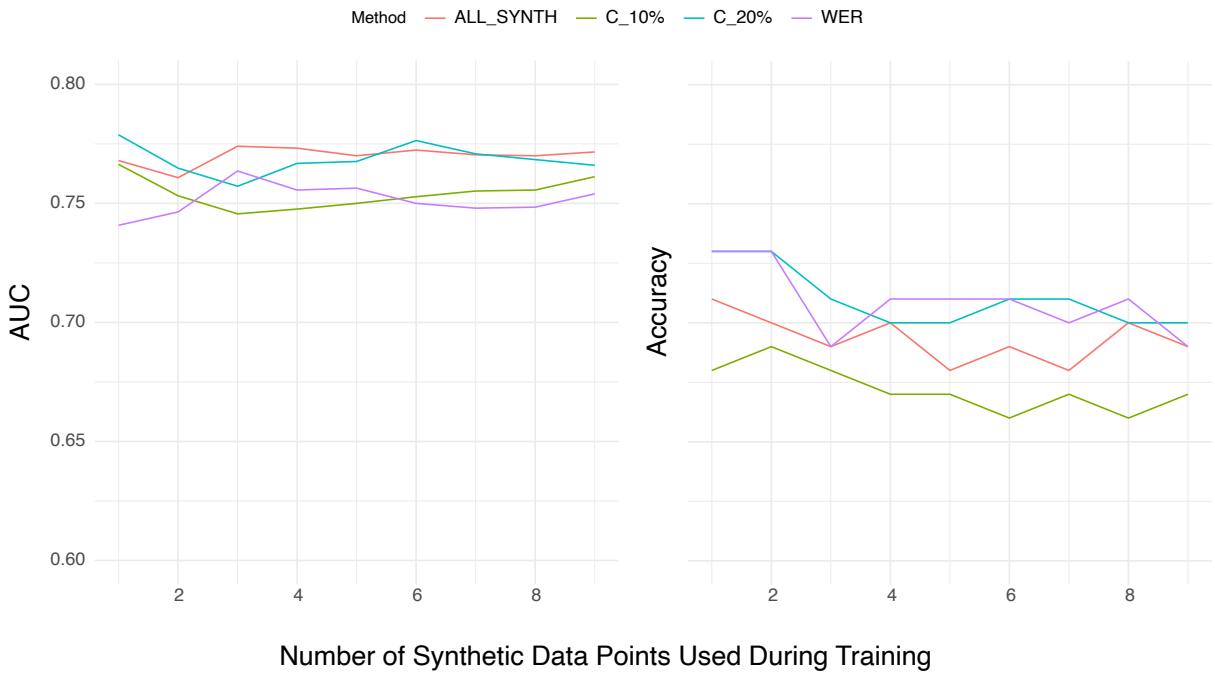

Figure 2: Visualization for the incremental synthetic data experiments. The number of synthatic data points used for training per participant is given on the x-axis. Area Under Curve (AUC).

N	AUC	Accuracy
1	0.76	0.71
2	0.76	0.71
3	0.76	0.69
4	0.76	0.70
5	0.76	0.69
6	0.76	0.69
7	0.76	0.69
8	0.76	0.69
9	0.76	0.69

Table 4: Summarization of incremental experiment results. N is the number of synthetic data points per participant used during training. AUC and accuracy are averaged over the augmentation method.

real data. One of the downsides to using WER is the need for expensive and time-consuming manual annotation. However, this can be bypassed with the constant rate method. In addition, constant deletion rates could be blended, similarly to what has been done with the ALL SYNTH method that achieved the highest AUC score for the SVM. Additional experiments determined that a linear model was not able to distinguish between the synthetic and real data based on the permutation test. Furthermore, as the amount of synthetic data used for training is increased consistent performance is reported that is comparable to the real data scenario.

One benefit of using random erasing to generate synthetic transcripts—rather than just simulating feature values from a distribution—is that the data is still explainable. There are synthetic transcripts that can be viewed and investigated. This is something that is highly sought after in clinical settings as medical professionals prefer tangible and explainable solutions.

These findings have an impact on future work. The ability to generate additional synthetic clinical data could open the door to training deep learning models and neural approaches. As well as, the raw data could be used for new solutions that are now possible due to increased amounts of data. For example, the sequence of words could be used as an input for an LSTM.

However, there are still some unknown factors of what this method has on data in other domains. For instance, this paper is scoped to a single clinical task that is focused on assessing cognition. It is unknown if this methodology would work on free speech clinical tasks, such as the picture description task or story telling task, where cognition and language abilities interact more heavily (Themistocleous et al., 2020). Future work would need to investigate the transference of this technique to other domains.

8. Conclusion

This paper proposed to generate synthetic data by simulating ASR error already found in automatic evaluation pipelines. Random erasing by either WER or constant deletion is a low cost and simple solution that effectively delivers machine learning performance that is on par with current real data methods. These findings present impactful solutions for future work to investigate how much data can be generated and achieving

better performance using deep learning and neural approaches.

9. Acknowledgements

This research was funded by MEPHESTO project Q10 (BMBF Grant Number 01IS20075).

10. Bibliography

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Clark, L., Gatz, M., Zheng, L., Chen, Y.-L., McCleary, C., and Mack, W. (2009). Longitudinal verbal fluency in normal aging, preclinical, and prevalent alzheimer's disease. *American journal of Alzheimer's disease and other dementias*, 24:461–8, 09.

Errattahi, R., El Hannani, A., and Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37. 1st International Conference on Natural Language and Speech Processing.

Farzanfar, D., Statucka, M., and Cohn, M. (2018). Automated indices of clustering and switching of semantic verbal fluency in parkinson's disease. *J Int Neuropsychol Soc*, 24(10):1047–1056, Nov.

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., and Robert, P. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and Geriatric Cognitive Disorders*, 45(3-4):198–209.

Konig, A., Lindsay, H., Tröger, J., and Ramakers, I. H. (2019). The use of artificial intelligence and automatic speech and image analysis for remote cognitive testing. In *Alzheimer Europe Conference, 23-25th October, The Hague, Netherlands.*, The Hague, Netherlands, October.

Lindsay, H., Müller, P., Linz, N., Zeghari, R., Magued Mina, M., Konig, A., and Tröger, J. (2021). Dissociating semantic and phonemic search strategies in the phonemic verbal fluency task in early dementia. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 32–44, Online, June. Association for Computational Linguistics.

Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., and Kokkinakis, D. (2019a). 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.

Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., and Kokkinakis, D. (2019b). 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.

McDonnell, M., Dill, L., Panos, S., Amano, S., Brown, W., Giurgius, S., Small, G., and Miller, K. (2020). Verbal fluency as a screening tool for mild cognitive impairment. *Int Psychogeriatr*, 32(9):1055–1062, Sep.

Pakhomov, S. and Hemmy, L. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55(1):97–106, June. Funding Information: The work on this study was supported in part by the National Institutes of Health National Library of Medicine Grant [LM00962301 – S.P.] and the Nun Study data collection was supported by a grant from the National Institute of Aging (R01AG09862). The authors also wish to thank Heather Hoecker for helping with digitization of the SVF samples.

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.

R Core Team, (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *Plos one*, 15(7):e0236009.

Tröger, J., Linz, N., König, A., Robert, P., and Alexandersson, J. (2018). Telephone-based dementia screening i: Automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '18, page 59–66, New York, NY, USA. Association for Computing Machinery.

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*, 131:53–61.

Vaughan, R., Coen, R., Kenny, R., and Lawlor, B. (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*, -1:1–7, 04.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2017). Random erasing data augmentation. *CoRR*, abs/1708.04896.