



Deep Learning Approaches for Detecting Alzheimer’s Dementia from Conversational Speech of ILSE Study

Ayimnisagul Ablimit*, Karen Scholz*, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany

ay.ablimit@uni-bremen.de

Abstract

Automatic screening of Alzheimer’s Dementia (AD) can have significant impact on society and the well-being of the patients. Early detection of AD from spontaneous speech offers great potential for inexpensive and convenient casual testing. We propose our deep neural network architecture that leverages acoustic, linguistic and demographic features to build a model for dementia screening for the biographic interview speech corpus of Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE). We oversample non-sequential and sequential features using well-known oversampling techniques and adapted data augmentation techniques to overcome the challenge of the imbalanced dataset, since the distribution of the diagnostic groups in ILSE corresponds to the prevalence of dementia. Our system achieves 70.6% of unweighted average recall on a 3-class classification problem. Moreover, we also investigate the feature importances to the model prediction to identify the most relevant indicators for AD detection, which may contribute to interpreting signs of cognitive decline and thus supporting clinicians in the diagnosis of dementia.

Index Terms: Speech & language, Alzheimer’s disease, acoustic and linguistic features, ILSE corpus, imbalanced data

1. Introduction

Dementia is an incurable progressive disease and the prevalence of dementia in people aged 65+ is estimated to be around 7% with prevalence increasing with age (World Health Organization report, 2021). The most common type of dementia is Alzheimer’s Dementia (AD). Although there is no curative therapeutic intervention for dementia, studies show that early interventions can delay the progression of the disease [1, 2]. Thus, it is important to recognize its symptoms as early as possible. Since speech and language ability can be affected at the early stage of cognitive deficits, spoken language skills are well established early indicators of cognitive deficits. In comparison to traditional diagnostic procedures, screening dementia and cognitive impairment from spontaneous speech provides a low-cost and widespread alternative for the assessment of cognitive states.

Studies have proposed methods to detect AD from speech and language features and achieved promising results [3, 4, 5, 6]. Several studies developed AD detection models for DementiaBank [7], a publicly available dataset, which contains audio recordings and corresponding transcripts of participants describing the Cookie Theft picture. Since the scenario in the picture description task is limited, studies used task specific features, for example *Information units* [8] detecting if participants

used specific words to capture main objects or actions of the picture, in addition to the acoustic and linguistic features [4]. However, it is challenging, because it requires linguistic and domain knowledge from experts and is restricted to tasks limited to specific topics. Moreover, varying characteristic between different stages of AD makes it harder for manual feature-engineering.

Our recent work [9, 10, 11] have focused on the detection of dementia from biographic interviews acquired within ILSE *Interdisciplinary Longitudinal Study on Adult Development and Aging* – a study which aims to investigate satisfying and healthy aging [12]. It consists of more than 6,500 hours of biographic interviews of over 1000 participants. Cognitive diagnoses of the patients in each time measurement are available. In our previous studies, more than 6000 acoustic and linguistic features were extracted for detecting AD and age associated cognitive decline (AACD). In contrast to DementiaBank, ILSE was not anticipated for speech based dementia screening. Diverse topics are covered by semi-structured biographic interviews, which results in difficulties in capturing grammatical or contextual patterns. Besides, severity of the AD or AACD varies from the patients and each time measurement. Furthermore, the features used for AD detection have been extracted on the interview level (varying from 2–6 hours), but temporal dynamics across the segments have not been investigated yet. These characteristics of ILSE make manual feature-engineering, effectively combining and capturing some specific patterns of these features from different feature categories for AD detection challenging.

Motivated by the shortcomings of manual feature-engineering and promising results of studies applied Convolutional Neural Network (CNN) - Long Short-Term Memory (LSTM) network based architecture, which efficiently combines features from multiple categories on AD detection of DementiaBank [13, 14, 15], we propose CNN-Gated Recurrent Unit (GRU) based architecture for 3-class cognitive impairment classification using sequential and interview-level features of recordings and transcripts from ILSE. In order to keep the balance of the dataset and the number of features, we apply random forest feature selection using hellinger distance as splitting metric, which is statistically appropriate for imbalanced datasets. For the purpose of overcoming the imbalanced dataset problem in ILSE, we implement weighted Dynamic Time Warping Barycentric Averaging (DBA), which was originally designed for data augmentation purposes [16], to oversample our sequential features. Moreover, with the aim of interpreting the indicators for AD detection, we further analyze the importance of the individual features to AD detection by calculating the derivative of the final prediction with respect to the input features.

2. Related work

Karlekar et al. (2018) [14] applied CNN-LSTM models using minimal feature engineering namely word embeddings and Part-of-Speech (POS) Tags as input and achieved a new Bench-

*Both authors contributed equally to this work. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project “ALMED - acoustic and linguistic features for early prediction of cognitive deficits” (403605461)

mark on DementiaBank. Mahajan et al. (2021) [15] explored deep neural network based linguistic and acoustic model and combined them into a multimodal model and applied it on DementiaBank ADReSS [17] dataset. By the linguistic model, similar to the work by Karlekar et al.(2018), features are fed into a CNN- bidirectional LSTM followed by self-attention and enriched the resulting representation with engineered feature including psycholinguistic, sentiment and demographic features. Extracted acoustic features on the audio segments are fed into GRU, which combines segment-level features into a common vector while maintaining the temporal structure across segments. Linguistic and acoustic models are combined with Fully Connected Neural Network (FCNN). They confirm the use of attention mechanism improves the performance and there exist temporal patterns in the linguistic model. However authors claim that the applied linear fusion method combining the different representations from different modality was too simple and suggest more complex combination methods.

3. The ILSE study

The ILSE study was initiated with the aim to investigate satisfying and healthy aging in middle adulthood and later life [18, 12]. Over the course of 20 years, more than 6500 hours of biographic interviews was collected from over 1,000 participants residing in east and west Germany in two cohorts. Each participant engaged in up to four measurements. It enables aging-related research in many disciplines including geriatrics, psychology, gerontology, sociology, history, and linguistics [18]. The biographic interviews were conducted in semi-standardized manner, where participants were asked to give detailed elaborations on the standardized open questions. The duration of the interviews became shorter across measurements, since the gathered biographic information accumulated over time. The speech recordings of the first two measurements were stored on tapes. From the third measurement, digital recording devices were used. All interviews have been digitized using a sampling rate of 16 kHz, and 16-bit linear PCM quantization. From the interviews of more than 1000 participants, as of today 145 interviews (about 440 hours) from 91 participants have been manually transcribed. These transcripts do not include speaker-turn wise time alignment information. Since it is an ongoing work, the number of interviews with manual transcripts is growing. In order to distinguish it from our other experiments, we denote this subset as $ILSE_{m145}$. Cognitive diagnoses of $ILSE_{m145}$ include: controls, AD and AACD. We refer to an interview with corresponding diagnostic label as one sample. The number of interviews and demographic information: age and year of education, is shown in Table 1.

4. Method

4.1. Feature extraction for screening

Although there is no speaker-turn wise time alignment in the manual transcripts, speaker-turn annotation is provided. Linguistic features are extracted directly from manual transcripts and before extracting acoustic features, speaker diarization [3] is applied to exclude speech from interviewers.

We extracted speech and linguistic indicators to capture speaker characteristics of how they speak and what they say. With the intention of analyzing the temporal pattern of sequences and classification experiments, we extracted two sets of features: interview-level and block-wise features. In the case of block-wise features, segments are accumulated with chronological order into a block of segments until the duration/number of

utterances reaches one minute (for acoustic features) or five utterances (for linguistic features), and then features are generated on this accumulated block-segments. Extracted indicators for this study include: Linguistic Inquiry and Word Count (LIWC) [19], Part-of-Speech (PoS) Tags [20] and PoS-perplexity [21], Perplexity [21], ivector [22], Voice activity detection (VAD) [3] based features and openSMILE [23]. Dimensions of the features are shown in Table 2. For a full description of the feature set and how they are generated, see Weiner et al. (2016) [3].

Table 1: *Demographic information per diagnostic class.*

M: Mean value, std: standard deviation

	control <i>M(std)</i>	AD <i>M(std)</i>	AACD <i>M(std)</i>
Interviews	108	16	21
Age	65.01 (4.02)	74.25 (0.44)	66.23 (2.52)
Year of education	13.46 (2.80)	11.43 (1.82)	12.38 (2.92)

4.2. Feature selection

The dataset is small regarding the large number of features. In order to keep the trainable weights of NNs low and prevent them from overfitting, we apply random forest classifiers based feature selection using hellinger distance as splitting metric, which is statistically appropriate for imbalanced datasets [24, 25]. Feature selection is applied on interview-level features of all feature sets respectively. Feature selection is conducted by leave-one-subject-out cross-validation and in each iteration 10 features are selected. For each feature, we keep the importance score and also accumulate the occurrences over the iterations. After the cross validation, for each feature, the mean value of importance scores during cross validation will be assigned as its importance score and the accumulated occurrences will be assigned as its occurrences. For the feature set, we calculate average feature importance and number of average occurrences over all the features. In the end, features with above average importance scores and occurrences will be selected. The number of selected features is shown in Table 2.

Table 2: *Feature dimension before and after feature selection*

	original dimension	selected features
LIWC	68	16
Perplexity	15	6
PoS	57	4
PoS-perplexity	23	9
lexical richness	2	2
ivector	128	8
VAD	12	8
openSMILE	6373	8

4.3. Feature categories and fusion

The extracted 7 sets of features are divided into five feature categories. We combine lexical richness and perplexity features, which are both related to vocabulary of the speaker. As second feature category, LIWC and PoS features are combined, which describe semantic and syntactic structure of speech. Each acoustic feature set is regarded as separate feature category.

As input for the classification task, LIWC & PoS features are considered as sequential features with the intention to capture varying contents and structures of sentences. OpenSMILE features are also considered as sequential features, in order to

capture varying characteristics of the speech. Other feature categories are applied as interview-level with following consideration:

- lexical richness and perplexity features: since vocabulary size is a speaker-related index, we assume that the estimation of vocabulary gets more precise when longer speech segments are considered.
- ivector: we assume ivectors remain nearly constant over time, since ivectors are originally used for speaker identification.
- VAD features are assumed to be more precise when longer segments are considered.

In order to verify our expectations, we calculate cosine distances between feature vectors of consecutive time steps, i.e., between consecutive one-minute speech segments for acoustic features and five utterances for linguistic features. First, features are normalized with zero-mean and unit-variance. Then consecutive cosine distances are calculated and averaged over all time steps and samples. As shown in Table 3, the results confirm our assumption: LIWC & POS feature category and openSMILE feature categories have high average cosine distance (around 0.8), which implies the variation between time steps. In contrast, other feature categories have low average cosine distance indicating less variation between time steps. It is worth mentioning that the obtained average cosine distance of the feature category lexical richness & perplexity is ca. 0.59. Nevertheless, we still persist that longer context yields more precise results for vocabulary based features.

Table 3: Average cosine distances between feature vectors of consecutive time steps per feature category

feature category	cosine distances
LIWC & PoS features	0.82
openSMILE	0.84
lexical richness & perplexity	0.59
VAD	0.37
ivector	0.19

4.4. Oversampling

The ILSE_{m145} is highly imbalanced with AD and AACD being underrepresented, which may result in poor classifier performance. Therefore, we apply oversampling which generates new samples for underrepresented classes to obtain a balanced dataset. We use the prominent Synthetic Minority Oversampling Technique SMOTE [26] to generate synthetic samples of interview-level features by regarding 5 nearest neighbours ($k = 5$). However, SMOTE is not applicable to sequential data. A naive approach to construct new samples of sequential features would be to average over features at each time step of the sequences. In that case temporal patterns are not taken into account and it may destroy temporal pattern of the features. Hence, we use weighted DBA, which was initially not developed as an oversampling method, but has been successfully applied for the purpose of data augmentation [16], for constructing synthetic samples for sequential features. Weighted DBA based new samples are constructed with following steps:

1. the k nearest neighbors of the sample under consideration (denoted as s) are identified using Dynamic Time Warping (DTW) distance as distance metric.
2. DTW alignment between s and each of its k -neighbors are calculated

3. New sample s_{new} is obtained by calculating the weighted average over time steps of s and time steps of its k -neighbors, which have been aligned by DTW.

As shown in Equation (1), $s_{new,t}$ denotes the constructed new sample at time step t , w_s and w_{n_i} denote the weight assigned to samples s and its neighbor n_i . DTW stands for the DTW function, which outputs the feature vector of n_i aligned with feature vector of time step t of s . We use dtw-python [27] to calculate DTW distances and alignments. We set the number of k -neighbors to 5. The sample under consideration (s) is assigned with a weight of 0.5 (w_s). Among the 5 nearest neighbors (n_i), two samples will be randomly selected and each assigned with weight of 0.15 (w_{n_i}) and the remaining weight of 0.2 will be equally divided to other 3 neighbors.

$$s_{new,t} = w_s \cdot s_t + \sum_{i=1}^k w_{n_i} \cdot DTW(s_t, n_i) \quad (1)$$

4.5. Network architecture

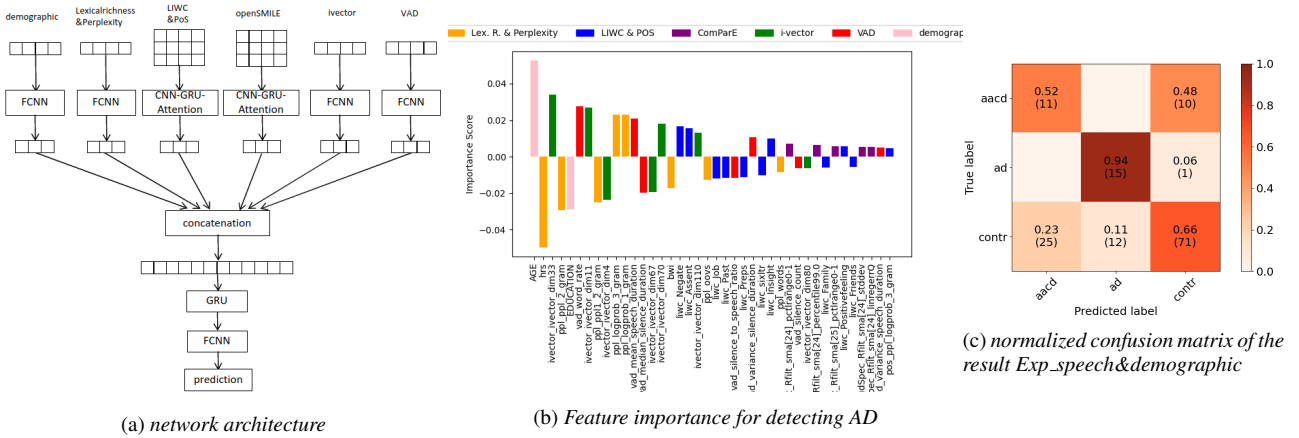
We use two different types of neural network structure: sequential neural networks for block-wise feature-categories as input, and non-sequential networks for interview-level feature-categories. The sequential NNs are adaptations of the CNN-LSTMs applied by Karlekar et al. (2018) [14]. We replace the LSTM by bidirectional GRU to reduce the number of trainable parameters and apply self-attention after GRU layer, which was proposed by Yang et al. (2016) [28]. The architecture of the network is shown in Figure 1a. Basically it consists of one-dimensional CNN followed by a bidirectional GRU. CNN is responsible for extracting local features within input sequences and GRU models long-term dependencies of the sequences. As input it receives the sequences of feature vectors from block-wise feature segments. The input data is processed by two single-layer CNNs using kernel sizes of 3 and 5, respectively and output of CNNs are concatenated. After applying dropout (dropout rate: 0.65) to the concatenated output, it feeds into bidirectional GRU layer and then self-attention is applied to the hidden states of bidirectional GRUs. The non-sequential NN consists of single-layer FCNN with ReLU activation. Interview-level features are fed into the FCNNs and hidden representation will be generated.

The feature vectors of five feature categories are processed by separate networks. LIWC & PoS and openSMILE features are processed by the proposed sequential NN. Lexical richness & perplexity, ivector and VAD features are processed by the non-sequential network. The resulting hidden representations are concatenated and fed into GRU, which provides more complex fusion possibility of features from different modalities. The fused representation is then passed to a single-layer FCNN, which outputs the final prediction. The network is trained for 60 epochs using Adam optimizer.

4.6. Classification

Random forest based selected features are fed into the network. For network training, we apply leave-one-subject-out cross-validation. The samples from the respective person are used as test set. The remaining samples are divided into training (90%) and validation set (10%). The samples of the validation are randomly selected so that samples of each diagnostic class come up in the validation set. Oversampling methods are applied on the training data. In our first set of experiment (Exp_onlySpeech), only the features from the five feature categories introduced in Section 4.1 are fed into the network. In our second set of experiment (Exp_speech&demographic), demographic information:

Figure 1



age, year of education is regarded as another feature category and also fed into the network along with existing feature categories in *Exp_onlySpeech*. We evaluate our results in terms of Unweighted Average Recall (UAR), since the dataset is imbalanced.

Using the selected interview-level features of these 7 feature sets, we train Support Vector Machines (SVM) based classifier as baseline (*Exp_baselineSVM*).

4.7. Feature importance

In order to further investigate the contribution of individual features to the trained model for AD detection, we calculate the feature importance by taking the advantage of weighting of the input features inside a NN. Our approach is based on the idea of saliency maps proposed by Simonyan et al. (2013) [29], who calculated saliency maps to visualize the impact of each pixel on the final prediction in an image classification task. We estimate the impact of the features to the predicted diagnose by calculating the derivative of the final prediction with respect to the input features. The obtained derivatives indicate in which direction and how much a feature value should have to be changed in order to obtain higher class score on the respective class. High magnitudes indicate small changes have a higher impact of the final prediction. Therefore, the gradient can be regarded as an estimator for the feature importance of the model. While feature importance of interview-level features are directly contained by calculating the derivatives, for the sequential features, we calculate the derivative of feature x at all time steps and assign the importance score of the feature with maximum magnitude across all time steps to the feature x . At the end, feature importances are calculated per class by averaging over feature importances of all correctly predicted samples of respective classes.

5. Results

Classification The baseline with SVM classifier (*Exp_baselineSVM*) achieved UAR of 62.6%. By using only acoustic and linguistic features (*Exp_onlySpeech*), we achieved UAR of 69.7%. The recall of AD is 87.5%. There are cases in which healthy controls are predicted as AACD and vice versa. As shown in Figure 1c, by adding demographic information in *Exp_speech&Demographic*, we achieved UAR of 70.6%, which is nearly 1% absolute improvement compared to using experiments without demographic information. Similar

to the case of first set of our experiments, misclassification between healthy controls and AACD occurred also by this model. On the one hand, the classifier may not be robust enough to differentiate these two classes from each other. On the other hand, healthy controls might be hard to distinguish from AACD in this dataset.

Feature importance In Figure 1b, we visualize the features according to their absolute feature importance score. A positive value of importance indicates positive correlation for predicting AD and vice versa. For simplicity we visualized only the 40 highest-ranked features. Demographic features, especially *age* is ranked as the most important feature and also positively associated with prediction of AD. As is known, the prevalence of AD increases with aging. Both linguistic and acoustic features are highly ranked. Highest ranked linguistic features are lexical richness (*hrs*) and perplexity features. The visualization confirms, vocabulary and speech complexity are negatively correlated with AD, more precisely, people with AD tend to have smaller vocabulary and use less diversity in expression. Ivectors and VAD features are highly ranked acoustic features. Besides confirming some common fact of dementia detection, this kind of visualization is helpful for the further analysis what the trained model learned.

6. Conclusions

This work proposed a deep neural network based approach leveraging acoustic, linguistic and demographic features to build an AD detection system. We also took several steps forward to overcome challenges of dementia detection task especially in ILSE. In particular, we analyzed the features to determine temporal variation of the sequence. We applied different oversampling methods on interview-level and sequential features respectively to overcome the imbalance of the dataset. Based on the trained model, we also calculate the feature importance score to interpret which indicators contribute more to the detection of AD. The purposed model surpass the baseline model trained with SVM classifier. Nevertheless, the model is not robust enough to discriminate AACD and healthy controls. Hence, it calls for the future work to explore the inside of the model and analyze mostly contributed features individually to the prediction of respective diagnostics.

7. References

- [1] M. J. Prince, A. Wimo, M. M. Guerchet, G. C. Ali, Y.-T. Wu, and M. Prina, "World alzheimer report 2015-the global impact of dementia: An analysis of prevalence, incidence, cost and trends," 2015.
- [2] D. Hsu and G. A. Marshall, "Primary and secondary prevention trials in alzheimer disease: looking back, moving forward," *Current Alzheimer Research*, vol. 14, no. 4, pp. 426–440, 2017.
- [3] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *Interspeech*, 2016, pp. 1938–1942.
- [4] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.
- [5] W. Kong, H. Jang, G. Carenini, and T. Field, "A neural model for predicting dementia from language," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 270–286.
- [6] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting ad," in *Proceedings of Interspeech 2019*. International Speech Communication Association (ISCA), 2019, pp. 4105–4109.
- [7] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.
- [8] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "Comparative study of oral and written picture description in patients with alzheimer's disease," *Brain and language*, vol. 53, no. 1, pp. 1–19, 1996.
- [9] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *ASRU*, 2019.
- [10] A. Abulimiti, J. Weiner, and T. Schultz, "Automatic Speech Recognition for ILSE-Interviews: Longitudinal Conversational Speech Recordings Covering Aging and Cognitive Decline," in *Proc. Interspeech 2020*, 2020, pp. 3795–3799.
- [11] A. Ablimit and T. Schultz, "Automatic speech recognition for dementia screening using ilse-interviews," in *Speech Communication; 14th ITG Conference*, 2021, pp. 1–5.
- [12] C. Sattler, H.-W. Wahl, J. Schröder, A. Kruse, P. Schönknecht, U. Kunzmann, and A. Zenthöfer, "Interdisciplinary longitudinal study on adult development and aging (ILSE)," *Encyclopedia of geropsychology*, pp. 1–10, 2015.
- [13] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," *arXiv preprint arXiv:1906.05483*, 2019.
- [14] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," *arXiv preprint arXiv:1804.06440*, 2018.
- [15] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech," *Frontiers in Aging Neuroscience*, p. 20, 2021.
- [16] G. Forestier, F. Petitjean, H. A. Dau, G. I. Webb, and E. Keogh, "Generating synthetic time series to augment sparse datasets," in *2017 IEEE international conference on data mining (ICDM)*. IEEE, 2017, pp. 865–870.
- [17] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.
- [18] P. Martin, M. Grünendahl, and M. Schmitt, "Persönlichkeit, kognitive leistungsfähigkeit und gesundheit in ost und west: Ergebnisse der interdisziplinären längsschnittstudie des erwachsenenalters (ilse)," *Zeitschrift für Gerontologie und Geriatrie*, vol. 33, no. 2, pp. 111–123, 2000.
- [19] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.
- [20] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *New methods in language processing*, 2013, p. 154.
- [21] C. Frankenberg *et al.*, "Perplexity – a new predictor of cognitive changes in spoken language? – results of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE)," *Linguistics Vanguard*, vol. 5, 06 2019, s2. [Online]. Available: <https://www.degruyter.com/view/j/lingvan.2019.5.issue-s2/lingvan-2018-0026/lingvan-2018-0026.xml>
- [22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [23] F. Eyben, F. Wengler, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.
- [24] R. Aler, J. M. Valls, and H. Boström, "Study of hellinger distance as a splitting metric for random forests in balanced and imbalanced classification datasets," *Expert Systems with Applications*, vol. 149, p. 113264, 2020.
- [25] G.-H. Fu, Y.-J. Wu, M.-J. Zong, and J. Pan, "Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data," *BMC bioinformatics*, vol. 21, no. 1, pp. 1–14, 2020.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [27] T. Giorgino, "Computing and visualizing dynamic time warping alignments in r: the dtw package," *Journal of statistical Software*, vol. 31, pp. 1–24, 2009.
- [28] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.
- [29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.