Interspeech 2024
1-5 September 2024, Kos, Greece

Translingual Language Markers for Cognitive Assessment from Spontaneous Speech

Bao Hoang[1], *Yijiang Pang*[1], *Hiroko Dodge*[2], *Jiayu Zhou*[1]

[1]Department of Computer Science and Engineering, Michigan State University, USA
[2]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, USA

hoangbao@msu.edu, pangyiji@msu.edu, hdodge@mgh.harvard.edu, jiayuz@msu.edu

Abstract

Mild Cognitive Impairment (MCI) is considered a prodromal stage of dementia, including Alzheimer's disease. It is characterized by behavioral changes and decreased cognitive function, while individuals can still maintain their independence. Early detection of MCI is critical, as it allows for timely intervention, enrichment of clinical trial cohorts, and the development of therapeutic approaches. Recently, language markers have been shown to be a promising approach to identifying MCI in a non-intrusive, affordable, and accessible fashion. In the InterSpeech 2024 TAUKADIAL Challenge, we study language markers from spontaneous speech in English and Chinese and use the bilingual language markers to identify MCI cases and predict the Mini-Mental Status Examination (MMSE) scores. Our proposed framework combines the power from 1) feature extraction of a comprehensive set of bilingual acoustic features, and semantic and syntactic features from language models; 2) careful treatment of model complexity for small sample size; 3) consideration of imbalanced demographic structure, potential outlier removal, and a multi-task treatment that uses the prediction of clinical classification as prior for MMSE prediction. The proposed approach delivers an average of 78.2% Balanced Accuracy in MCI detection and an averaged RMSE of 2.705 in predicting MMSE. Our empirical evaluation shows that translingual language markers can improve the detection of MCI from spontaneous speech. Our codes are provided in https://github.com/illidanlab/translingual-language-markers.

Index Terms: Mild Cognitive Impairment Detection, Translingual Language Markers, Computational Paralinguistics

1. Introduction

Alzheimer's disease (AD) is a type of dementia that impacts memory, cognition, and behavior and ranks as the seventh-leading cause of death in the United States in 2020 [1]. Mild Cognitive Impairment (MCI) is the prodromal stage of dementia, including AD, characterized by minor problems with memory loss, speech and language impairment, and reasoning difficulties. Early detection of MCI is critical, allowing for timely intervention and improvements in quality of life and enabling cohort enrichment towards the understanding of pathology and the development of therapeutical approaches.

Even though *in vivo* markers and imaging markers from brain scans are shown to be very sensitive in the detection of MCI [2], they are not easily accessible nor generally affordable for screening. Recently, language markers have been shown to be a promising approach to identifying MCI in a non-intrusive, affordable, and accessible fashion. The effectiveness of language markers is studied in the context of semi-structured conversation [3, 4, 5] and spontaneous speech [6], and showed promising predictive power differentiating MCI and cognitive normal subjects. In the InterSpeech 2024 TAUKADIAL Challenge, we study language markers from spontaneous speech in English and Chinese and use the language markers to identify MCI cases and predict the Mini-Mental Status Examination (MMSE) scores. There are many outstanding challenges. The first one is the small sample size: we have only 62 English speakers and 64 Chinese speakers in the provided training data. Building predictive models separately for the two languages greatly limits the number of markers that can be explored and included in the model due to the restricted model complexity needed to prevent overfitting. How to jointly consider all samples in a unified predictive pipeline is critical to ensure the prediction performance.

In this paper, we conduct extensive experiments and propose a cross-lingual strategy to combine the information from the two languages. Our proposed framework extracts a comprehensive set of features, including acoustic features based on the raw speech, and embedding from pre-trained language models that capture interactions among semantic and syntactic elements in transcribed text. We used machine neural translation from Chinese to English to secure a set of shared embedding features with English and applied back-translation in English to remove the impact of bias induced by the translation system. To control model complexity, we select the most relevant features to be included in the model by ranking the features using a supervised sparse learning model. To further improve the prediction performance, we explore the demographic structure, identify imbalance subgroups that may induce undesired bias in the models, and finally remove them by constructing a weighted loss function. We incorporate a two-staged procedure that identifies samples with potentially noisy labels and eliminates them in the final learning. Finally, we develop a multi-task treatment that couples the two prediction tasks, using the prediction of clinical classification as prior for MMSE prediction. The proposed approach delivers an average of 78.2% Balanced Accuracy in MCI detection and an averaged RMSE of 2.705 in predicting MMSE. Our empirical evaluation shows that translingual language markers can improve the detection of MCI from spontaneous speech.

2. Dataset

The TAUKADIAL Challenge Dataset [7] consists of spontaneous speech samples corresponding to audio recordings of picture description tasks produced by both cognitively normal subjects and patients diagnosed with MCI. Each subject has three individual audio recordings corresponding to descriptions of three corresponding pictures. The dataset has been evenly balanced with respect to age and gender to eliminate poten-

10.21437/Interspeech.2024-1422

Language	English (En)		Chinese (Zh)	
Cognitive status	NC	MCI	NC	MCI
Number of subjects	21	41	34	33
MMSE	29.2±0.6	27.8±1.2	29.0±1.3	23.5±4.3
Gender (%female)	85.7%	56.1%	47.1%	66.7%
Age	69.7±6.2	72.0±6.6	73.2±6.5	75.1±4.9

Table 1: *Demographics of training dataset in the TAUKADIAL Challenge. The language labels (En and Zh) are identified by automatic speech recognition.*

tial confounding and bias [7]. The dataset includes both English (En) and Chinese (Zh) speakers and the language used by each participant is identified by an Automatic Speech Recognition (ASR) model with details later. The three pictures used for English and Chinese speakers are different. The training dataset has 129 participants in total, including 67 participants who identified as Chinese speakers and 62 as English speakers. A clinical classification is provided for each subject, either healthy normal cognition (NC) or mild cognitive impairment (MCI). Among the 129 participants, we have 74 MCI patients, of which 33 are Chinese speakers, and 41 are English speakers. Among the 55 cognitively healthy individuals, there are 34 Chinese speakers, and 21 are English speakers. Mini-Mental Status Examination (MMSE) score is provided for each training subject as a target variable for the regression task. We summarize the demographic information in Table 1. The test dataset includes 40 subjects, and the goals of the competition are to 1) predict the clinical classification (NC/MCI) and 2) predict the MMSE score of each subject.

3. Methodology

For the two tasks, clinical classification and MMSE score prediction, we first extract various types of features from the audio, including acoustic features and language features, and then conduct predictive modeling with feature selection. One technical challenge is that the modeling includes two different languages, and our proposed approach jointly considers and aligns two languages to improve predictive performance. We have also developed strategies to mitigate imbalanced classes and outlier detection. Figure 1 overviews the technical components of the proposed approach.

3.1. Feature Extraction

3.1.1. Acoustic Features

Acoustic features have been shown to include information identifying cognitive impairments [6, 4]. We use Python Library librosa [8] and open-source software for features extraction from audio signals OpenSMILE [9] for audio preprocessing and acoustic feature extraction. We obtained Mel-frequency cepstral coefficients (MFCCs) from one speech following the method described in [10]. This includes extracting the first 13 MFCC bands (0-12) along with their corresponding 13 delta MFCCs and 13 delta-delta MFCCs, representing the rate of change and acceleration in MFCCs. After that, we apply 6 descriptive statistics functions (mean, standard deviation, variance, max, min, median), generating a total of 234-dimensional MFCC features for one speech. In addition, we also obtain the extended version of Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [11], which contains 18 Low-level descriptors (LLD) from original GeMAPS, encompassing frequency-related parameters such as pitch, jitter, formant, as well as energy-related parameters like shimmer, loudness, and harmonics-to-noise ratio (HNR), along with various spectral parameters, and incorporates an additional 7 LLDs, including MFCC 1-4, Spectral flux, and Formant 2-3 bandwidth. After that, we apply statistical functions to each LLD, resulting in 88-dimensional eGeMAPS features for one speech.

Since each subject has three speech recordings, one for each picture, we independently extract and then concatenate these MFCC features of all three speeches on a specific subject.

3.1.2. Semantic and Syntactic Features

Semantic and syntactic information in speech has previously been shown to be informative in detecting early dementia [3, 6, 4]. With a limited training sample size, we rely on pre-trained language embeddings to capture such semantic and syntactic information. Pre-trained by a large-scale public corpus, a language embedding outputs a fixed-length numerical vector given a word or sentence and captures semantic and syntactic relationships. Meanwhile, in order to maximize the utility of data, we propose to develop a novel translinguistic embedding that extracts the same set of features for speakers from two languages.

Step 1: Audio transcription. We use OpenAI's Automatic Speech Recognition (ASR) model Whisper [12] to transcript the raw audio files. Whisper is trained on a vast dataset of 680,000 hours of multilingual and multitask supervised data obtained from the web and supports multiple language tasks, including both English and Chinese ASR. It lets us simultaneously detect the language and transcribe speech audio files from English and Chinese into text.

Step 2: Cross-language alignment. In order to perform a joint analysis of two languages and greatly improve the sample size used by our prediction models, a straightforward strategy is to translate one language to another and extract embedding features using the same language model. However, the translation induces biases from the translation process and such biases will cause additional distributional differences that compromises the prediction performance for such joint modeling, which is also empirically validated through our experiments. To this end, we propose adopting a back-translation strategy for alignment, a common data augmentation strategy in machine translation [13]. In the process of translating transcripts from one language to another, we utilized Facebook's M2M100 multilingual sequence-to-sequence model [14], designed to facilitate translation across 100 languages. We translate Chinese transcripts into English and extract embedding from the translated English text. For English transcripts, instead of directly using them for embedding extraction, we first translate the English transcripts into Chinese and then back-translate them back into English. This two-step translation method *aligns* distributional differences induced by the machine translation process and is expected to maintain the semantic information.

Step 3: Computing Embeddings Given the aligned transcriptions, we use pre-trained Deep Bidirectional Encoder Representations from Transformers (BERT) [15] through the Huggingface Transformers library [16] to compute embedding features from the transcriptions. Specifically, we aggregate the embeddings of all words of one speech by taking the mean to generate a 768-dimensional transcript-level representation. We then concatenate three embedding features of all three speeches on a specific subject.

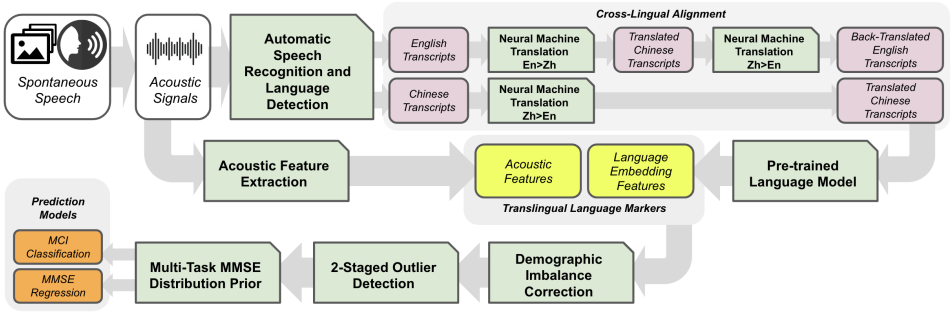

Figure 1: *Overview of the proposed approach.*

3.2. Feature Selection and Prediction Models

Since the dimension of the extracted features is substantially larger than the number of samples in our training data, directly using the features for prediction models is very likely to overfit and lead to poor prediction performance. To address this issue, we use feature selection to select the most relevant features and use them in modeling. In our final solution, we utilized the supervised feature selection based on sparse learning technique LASSO [17], and we use the implementation from the Scikit-learn library [18]. Specifically, we extract the top-k largest coefficients in absolute value from the LASSO model and use them as the k features for subsequent use in classification and regression tasks, where the hyper-parameter k is chosen based on the validation data. For classification, we choose the linear model due to the small sample size. We select the Logistic Regression (LR) from Scikit-learn library [18] and a customized PyTorch-implemented LR as the default classifiers for cognitive classification tasks. Specifically, the customized LR will conveniently support special considerations that are proposed to fit the data patterns. We also selected Support Vector Machine for Regression (SVR) and Random Forest for MMSE score regression task. For SVR, we use the validation data to choose the best-performing hyperparameters (RBF kernel with C = 8).

During model development, we randomly split the Competition training data into training and validation, and we can use this internal validation scheme to evaluate the performance and stability of models and choose model performance. We apply the best-performing models and hyper-parameters on the entire training data to generate the final submission.

3.3. Additional Treatments

Re-weighting imbalanced classes. From Table 1, we see that the number of subjects in main classes NC&MCI and the subclasses En&NC, En&MCI, Zh&NC, Zh&MCI are imbalanced. Such imbalance induces undesired bias in predictive models and compromises their generalization performance [19]. Therefore, we propose fixed weights to balance the training loss of each data point and incorporate it into one of our pipelines of classification tasks. Specifically, for each group (En&NC, En&MCI, Zh&NC, and Zh&MCI), we re-weight the data points with {0.4, 0.2, 0.3, and 0.3}, i.e., applying the corresponding weights to each loss value where the corresponding data points come from.

Outlier detection. Cognitive scores and clinical labels are notorious for their instability (e.g., discussions in [20, 21]). The potential existence of noisy labels complicates with the small sample size issue, leading to models of poor performance, as also evidenced by our empirical observation of unstable performance across multiple random seeds. To this end, we propose a two-phase training strategy to filter out the potential outliers in the whole training dataset. We first count the number of failure cases with wrongly high confidence during the first step of training, then filter out the subjects that appear multiple times for the second-step re-training.

Multi-tasking with MMSE distribution priors. From Table 1, we see that the MMSE ranges in the sub-classes En&NC, En&MCI, Zh&NC, Zh&MCI are different. Specifically, the low MMSE score basically only appears in the Zh&MCI category. As the clinical classification is generally an easier task than the regression task of MMSE prediction, we propose a multi-task strategy that considers the prediction of clinical label as prior knowledge, and incorporate it into one of our pipeline of regression tasks. Specifically, we first derive the language label and cognitive status of subjects, and then, individual regression models for each sub-class are trained and employed in MMSE score prediction.

4. Experiment

In this section, we evaluate the method's performance by randomly splitting the TAUKADIAL Challenge training data into 90% for training and 10% for validation. We repeat the experiment across 100 different random seeds, and we calculate the average and standard deviation of Balanced Accuracy and F1 for classification tasks, as well as Root Mean Square Error for regression tasks on the validation data. For test data performance, we report our 5 best hyperparameters performance.

4.1. Verification of Acoustic Features and Language Embeddings Features

Acoustic Features. We evaluate the performance of acoustic features and their combination. The averaged performance of 100 random seeds in Table 2 shows that MFCC features yield better performance in both classification and regression, attaining a balanced accuracy of 71.4% and a root mean square error of 2.773.

Feature Set	Balanced Accuracy	F1	RMSE
MFCC	**71.4±12.3**	**69.4±13.1**	**2.773±1.012**
eGeMAPS	60.4±12.8	57.1±13.8	2.949±1.165
MFCC + eGeMAPS	69.3±12.5	66.5±14.3	2.801±1.090

Table 2: *Performance of Acoustic Features*

Embedding Features. We compare the performance of two approaches towards multi-lingual language embedding: BERT

Feature Set	Balanced Accuracy	F1	RMSE
BERT[1]	**62.6±11.9**	**57.5±15.9**	2.935±1.039
Multilingual BERT	61.9±12.1	56.0±16.3	**2.837±1.065**

Table 3: *Performance of Language Embedding.* [1] *means back-translation method.*

Language Embeddings Acoustic Features	Balanced Accuracy	F1	RMSE
BERT[1] MFCC	**71.8±11.4**	**68.6±13.4**	**2.736±1.000**
Multilingual BERT MFCC	70.3±11.4	66.1±14.5	2.766±1.018

Table 4: *Performance of Acoustic Features + Language Embeddings.* [1] *means back-translation method.*

Number of top features	Balanced Accuracy	F1	RMSE
500	73.1±12.9	69.6±16.1	**2.648±0.882**
1000	74.2±12.6	70.8±15.6	2.663±0.900
1500	75.3±13.0	72.1±16.0	2.665±0.908
1600	**75.5±11.1**	**72.5±13.5**	2.714±0.928

Table 5: *Performance of LASSO Feature Selection*

with the proposed translation and Multilingual BERT by inputting two languages directly. The averaged performance in Table 3 shows that the original BERT with two-step translation achieves slightly better performance in the classification task with a balanced accuracy of 62.6%. On the other hand, Multilingual BERT exhibits better performance in the regression task, achieving a root mean square error of 2.837.

Combination of Acoustic Features and Embedding Features. We evaluate the combination of different types of acoustic features and language embedding features to find out the optimal overall performance. Table 4 shows that the best performance is achieved by the original BERT with two-step translation and MFCC acoustic features, with a balanced accuracy of 71.8% in the classification task and a root mean square error of 2.736 in the regression task.

4.2. Effects of Feature Selection Method

In our setting, the feature dimension is significantly larger than the sample size. Specifically, the combination of BERT embedding and MFCC acoustic features results in 3006 dimensions, whereas we only have 129 samples. We will investigate the efficacy of Lasso feature selection on the current optimal feature sets: the original BERT embedding from the two-step translation method combined with MFCC acoustic features, using varying numbers of top features. The results are in Table 5. Notably, the Lasso feature selection method enhances performance, particularly with 1600 top features achieving a balanced accuracy of 75.5% and 500 top features achieving a root mean square error of 2.648.

4.3. Verification of Additional Treatment

We verify the effectiveness of the proposed strategies in Section 3.3. All the conducted experiments use the same settings BERT embedding&MFCC features with Lasso feature selection (1600 features) under LR and SVR models. The results in Table 6 show that *Re-weighting imbalanced classes* actually has slightly decreased classification performance in the validation data. We expect performance improvement if the test data is more balanced. We see that the strategy of *Outlier detection* significantly improves averaged performance after filtering out some subjects with possible label issues. Finally, we see that the *multi-task prior* slightly improved the regression performance.

Setting	Balanced Accuracy	F1	RMSE
Default	75.5±11.1	72.5±13.5	2.714±0.928
Re-weighting	74.2±12.2	71.4±14.3	2.714±0.928
Outlier detection	**78.2±11.7**	**74.4±15.1**	2.719±0.901
Multi-task prior	75.5±11.1	72.5±13.5	**2.705±0.856**

Table 6: *Performance of Additional Treatments.*

Attempt	Balanced Accuracy	RMSE
1	**45.1**	3.095
2	42.5	2.928
3	42.0	3.220
4	42.0	**2.578**
5	43.9	2.732

Table 7: *Test Dataset Performance*

4.4. Test Dataset Performance

We select our top 5 hyperparameters based on an average of 100 random seeds of Balanced Accuracy and RMSE for validation data to submit test predictions to the TAUKADIAL Challenge Organizers. Hyperparameters include the number of top features, types of classifiers and regressors, different combinations of language embeddings and acoustic features, as well as combinations of additional treatments. Table 7 reports the results of our five attempts. Our proposed method achieves 45.1% balanced accuracy and 2.578 root mean square error.

5. Discussion and Conclusion

We evaluated the effectiveness of acoustic features and language embedding in detecting Mild Cognitive Impairment (MCI) and predicting Mini-Mental State Examination (MMSE) scores in a multilingual dataset. Upon comparing acoustic and language features, it was found that the acoustic-based model outperforms the language-based model. However, when both language embeddings and the acoustic model are utilized, the performance slightly increases to 71.8% in the classification task and a 2.736 RMSE in the regression task. Subsequently, applying a feature selection method to select the top 1600 features to prevent overfitting further improves the performance to 75.5% accuracy and a 2.714 RMSE. Finally, incorporating *Outlier detection* with *multi-task prior* results in the best classification performance of 78.2% balanced accuracy and a regression performance of 2.705 root mean square error.

In addition to the original BERT model, we also verified the language embeddings obtained from Google's Multilingual BERT, which supports various language inputs. This makes it more convenient to obtain the embeddings for English and Chinese transcriptions. However, the original BERT model, when combined with a two-step translation process and MFCC acoustic features, still outperforms Multilingual BERT with MFCC acoustic features.

For future directions, we intend to apply large language models such as GPT to investigate whether they can discern patterns between Mild Cognitive Impairment (MCI) and Normal Cognition (NC) among both English and Chinese speakers.

6. Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-2212174, IIS-1749940, Office of Naval Research N00014-20-1-2382, and National Institute on Aging (NIA) RF1AG072449, R01AG051628, R01AG056102.

7. References

[1] S. L. Murphy, K. D. Kochanek, J. Xu, and E. Arias, "Mortality in the united states, 2020," 2021.

[2] J. Zhou, J. Liu, V. A. Narayan, J. Ye, A. D. N. Initiative *et al.*, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, 2013.

[3] F. Tang, I. Uchendu, F. Wang, H. H. Dodge, and J. Zhou, "Scalable diagnostic screening of mild cognitive impairment using ai dialogue agent," *Scientific reports*, vol. 10, no. 1, p. 5732, 2020.

[4] F. Tang, J. Chen, H. H. Dodge, and J. Zhou, "The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment," *Frontiers in digital health*, vol. 3, p. 702772, 2022.

[5] B. Hoang, Y. Pang, H. H. Dodge, and J. Zhou, "Subject harmonization of digital biomarkers: Improved detection of mild cognitive impairment from language markers," in *Biocomputing 2024*. WORLD SCIENTIFIC, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1142/9789811286421_0015

[6] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of alzheimer's disease using spontaneous speech only," in *Interspeech*, vol. 2021. NIH Public Access, 2021, p. 3830.

[7] S. Luz, S. d. l. F. Garcia, F. Haider, D. Fromm, B. MacWhinney, A. Lanzi, Y.-N. Chang, C.-J. Chou, and Y.-C. Liu, "Connected speech-based cognitive assessment in chinese and english," 2024, final DOI to be assigned.

[8] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the Python in Science Conference*, ser. SciPy. SciPy, 2015. [Online]. Available: http://dx.doi.org/10.25080/Majora-7b98e3ed-003

[9] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, ser. MM '10. ACM, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1145/1873951.1874246

[10] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 1, 2001, pp. 73–76 vol.1.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, p. 190–202, Apr. 2016. [Online]. Available: http://dx.doi.org/10.1109/TAFFC.2015.2457417

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[13] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 86–96.

[14] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," 2020. [Online]. Available: https://arxiv.org/abs/2010.11125

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2019. [Online]. Available: https://arxiv.org/abs/1910.03771

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[18] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[19] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, J. Ye, A. D. N. Initiative *et al.*, "Analysis of sampling techniques for imbalanced data: An n= 648 adni study," *NeuroImage*, vol. 87, pp. 220–241, 2014.

[20] S. T. Pendlebury, F. C. Cuthbertson, S. J. Welch, Z. Mehta, and P. M. Rothwell, "Underestimation of cognitive impairment by mini-mental state examination versus the montreal cognitive assessment in patients with transient ischemic attack and stroke: a population-based study," *Stroke*, vol. 41, no. 6, pp. 1290–1293, 2010.

[21] C. Hörnsten, H. Littbrand, G. Boström, E. Rosendahl, L. Lundin-Olsson, P. Nordström, Y. Gustafson, and H. Lövheim, "Measurement error of the mini-mental state examination among individuals with dementia that reside in nursing homes," *European journal of ageing*, vol. 18, pp. 109–115, 2021.