Speech-Based Cognitive Impairment Classification Using Task-Specific Transformer Representations and Feature Fusion

Roneel V. Sharan, *Senior Member, IEEE*, Junhua Li, *Senior Member, IEEE*, and Minji Lee, *Member, IEEE*

***Abstract*— Early detection of cognitive impairment is important for timely clinical intervention and disease management. Speech and language have emerged as promising non-invasive biomarkers of cognitive decline, enabling automated assessment through natural language processing techniques. In this work, we propose a transformer-based framework for cognitive impairment detection using speech transcripts from multiple cognitive assessment tasks. Experiments are conducted on the PROCESS-2 dataset, which contains recordings from the Semantic Fluency Task (SFT), Phonemic Fluency Task (PFT), and Cookie Theft Description (CTD) task. Participants diagnosed with mild cognitive impairment and dementia are combined into a cognitively impaired group and distinguished from healthy controls. Separate BERT models are fine-tuned for each cognitive task to learn task-specific linguistic representations. Transcript-level embeddings are fused through feature concatenation and classified using a multilayer perceptron. Experimental results demonstrate that the BERT-based models outperform traditional machine learning approaches across all individual tasks. Among the single-task models, the CTD task achieved the highest F1-score of 84.71%. By integrating embeddings from all three cognitive tasks, the proposed fusion framework achieved an F1-score of 87.06%, surpassing the best individual-task model.**

I. INTRODUCTION

Cognitive impairment exists on a continuum ranging from mild cognitive impairment (MCI) to dementia and is associated with progressive decline in memory, language, executive function, and daily living abilities. MCI is often considered an intermediate stage between healthy cognitive aging and dementia, with a substantial proportion of affected individuals progressing to dementia over time [1], [2]. As global populations continue to age, the prevalence of cognitive impairment is expected to increase substantially, creating significant challenges for healthcare systems, caregivers, and society. Early identification of cognitive decline is therefore important for timely clinical intervention, patient management, and monitoring of disease progression [3], [4].

Current approaches for assessing cognitive impairment commonly involve neuropsychological testing, neuroimaging, and biomarker analysis [5]. While these methods provide valuable clinical information, they often require specialized equipment, trained personnel, and substantial financial resources. Furthermore, some procedures are invasive or difficult to deploy at scale, limiting their suitability for routine screening and longitudinal monitoring. Consequently, there is growing interest in developing accessible, low-cost, and non-invasive approaches for cognitive assessment.

Speech and language have emerged as promising digital biomarkers of cognitive health. Cognitive impairment is frequently associated with linguistic changes, including reduced lexical diversity, word-finding difficulties, semantic errors, impaired discourse organization, and alterations in verbal fluency [6]. These deficits can manifest during spontaneous speech and structured cognitive tasks, making speech-based assessment an attractive alternative to conventional diagnostic procedures. Advances in automatic speech processing and natural language processing (NLP) have further enabled the development of computational methods capable of identifying subtle language patterns associated with cognitive decline.

Recent studies have demonstrated the effectiveness of transformer-based language models for cognitive assessment from speech transcripts. In particular, large language models (LLMs) have generally outperformed traditional machine learning approaches based on hand-crafted linguistic features and bag-of-words representations [7], [8], [9]. The recently introduced PROCESS-2 dataset provides a large-scale benchmark for speech-based cognitive assessment and includes recordings from three complementary cognitive tasks: the Semantic Fluency Task (SFT), Phonemic Fluency Task (PFT), and Cookie Theft Description (CTD) task [9]. Benchmark experiments on PROCESS-2 showed that linguistic representations derived from speech transcripts are highly informative for distinguishing cognitively impaired individuals from healthy controls and that transformer-based models achieve superior performance compared with conventional approaches.

Although prior studies have demonstrated the effectiveness of language models on individual cognitive tasks, each task captures different aspects of cognitive functioning. SFT primarily evaluates semantic memory and lexical retrieval, PFT probes phonological retrieval and executive control, whereas CTD assesses connected speech production and discourse organization. Consequently, information obtained from these tasks may be complementary, suggesting that integrating task-specific representations could provide a more comprehensive characterization of cognitive status than relying on any individual task alone.

Motivated by this observation, this paper proposes a

This work was supported by the University of Essex. *(Corresponding author: Roneel V. Sharan.)*

Roneel V. Sharan and Junhua Li are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom (e-mail: roneel.sharan@essex.ac.uk; junhua.li@essex.ac.uk).

Minji Lee is with the Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon 14662, South Korea (e-mail: minjilee@catholic.ac.kr).

transformer-based framework for cognitive impairment detection using transcripts from multiple cognitive speech tasks. Separate BERT models are fine-tuned for SFT, PFT, and CTD transcripts to learn task-specific linguistic representations. The resulting embeddings are subsequently combined through feature-level fusion and classified using a multilayer perceptron (MLP). The proposed framework addresses a binary classification problem in which participants with MCI and dementia are grouped into a cognitively impaired (CI) class and distinguished from healthy controls (HC). Experimental evaluation on the PROCESS-2 dataset demonstrates that the proposed fusion strategy improves performance over individual task-specific models and conventional machine learning baselines.

II. MATERIALS AND METHODS

Fig. 1 illustrates the overall architecture of the proposed framework for cognitive impairment detection from speech transcripts. Three cognitive speech tasks, namely SFT, PFT, and CTD, are processed independently using task-specific fine-tuned BERT models to learn semantic representations. The resulting task-level embeddings are subsequently fused and classified using a MLP to distinguish CI from HC.

A. Dataset

Experiments were conducted using the PROCESS-2 dataset [9], a recently introduced speech corpus designed for automatic cognitive impairment assessment under real-world conditions. The dataset comprises speech recordings collected from 400 older adults, including participants diagnosed with dementia, individuals with MCI, and HC. Unlike laboratory controlled speech corpora, PROCESS-2 was acquired remotely using heterogeneous consumer devices and natural recording environments, thereby providing a realistic benchmark for developing clinically deployable speech-based cognitive screening systems.

Each participant completed three cognitive speech tasks: SFT [10], PFT [11], and CTD [12]. The fluency tasks assess lexical retrieval and executive functioning by requiring participants to generate words within a semantic or phonemic category, whereas the CTD task evaluates spontaneous connected speech through picture description. For every recording, manually generated transcripts are provided together with participant metadata, including age, gender, diagnosis, and cognitive assessment scores where available.

Following the predefined partitioning provided with the dataset, the corpus was divided into training and testing subsets containing 320 and 80 participants, respectively. The training split comprised 40 dementia, 120 MCI, and 160 HC participants, while the test split consisted of 10 dementia, 30 MCI, and 40 HC participants. Demographic statistics for each subset are summarized in Table I. Consistent with the findings reported for PROCESS-2, age distributions were broadly comparable across diagnostic groups, reducing the likelihood of age-related confounding effects during model development.

TABLE I
DEMOGRAPHIC CHARACTERISTICS OF THE PROCESS-2 DATASET USED IN THIS STUDY

Split	Diagnosis	*N*	Age (Years)	Gender (M/F)
Train	Dementia	40	75.00±8.15	28/12
	MCI	120	71.30±8.82	59/61
	HC	160	72.21±6.92	74/86
Test	Dementia	10	70.30±7.65	7/3
	MCI	30	70.47±8.91	15/15
	HC	40	73.98±6.94	19/21
Total	Dementia	50	74.06±8.20	35/15
	MCI	150	71.13±8.82	74/76
	HC	200	72.57±6.94	93/107
All	All	400	72.22±7.89	202/198

In this study, we formulated a binary cognitive impairment detection task by merging the MCI and dementia groups into a single cognitively impaired (CI) category. Consequently, the dataset contained 200 HC participants and 200 CI participants, yielding a balanced classification problem. This formulation is clinically relevant for large-scale screening applications, where the primary objective is to distinguish cognitively healthy individuals from those exhibiting signs of cognitive decline.

B. Task-Specific BERT Encoder

The proposed framework employs task-specific Bidirectional Encoder Representations from Transformers (BERT) models [13] to learn semantic representations from speech transcripts associated with the three cognitive assessment tasks. Given the relatively limited size of the PROCESS-2 corpus, a pretrained Tiny BERT model was adopted to reduce the number of trainable parameters during fine-tuning while retaining the contextual language modeling capabilities of transformer-based architectures. This design choice aims to mitigate overfitting and improve training efficiency on the available dataset. As illustrated in Fig. 1, transcripts obtained from the SFT, PFT, and CTD task are processed independently using separate fine-tuned BERT encoders. This design enables each model to learn task-dependent linguistic characteristics that may be indicative of cognitive impairment.

Prior to model training, each transcript is tokenized using the tokenizer associated with the pretrained BERT model. The tokenized sequences are then passed through the pretrained transformer encoder, which generates contextualized token representations by modeling bidirectional dependencies within the transcript. The BERT architecture employed in this work produces a 128-dimensional hidden representation for each token at its final encoder layer.

To obtain a fixed-length representation of the entire transcript, a Global Average Pooling (GAP) layer is applied to the final token embeddings generated by the transformer. Let $H = h_1, h_2, \ldots, h_T$ denote the sequence of token embeddings produced by the encoder, where $h_t \in \mathbb{R}^{128}$ and T represents the number of tokens. The transcript-level embedding z is computed as

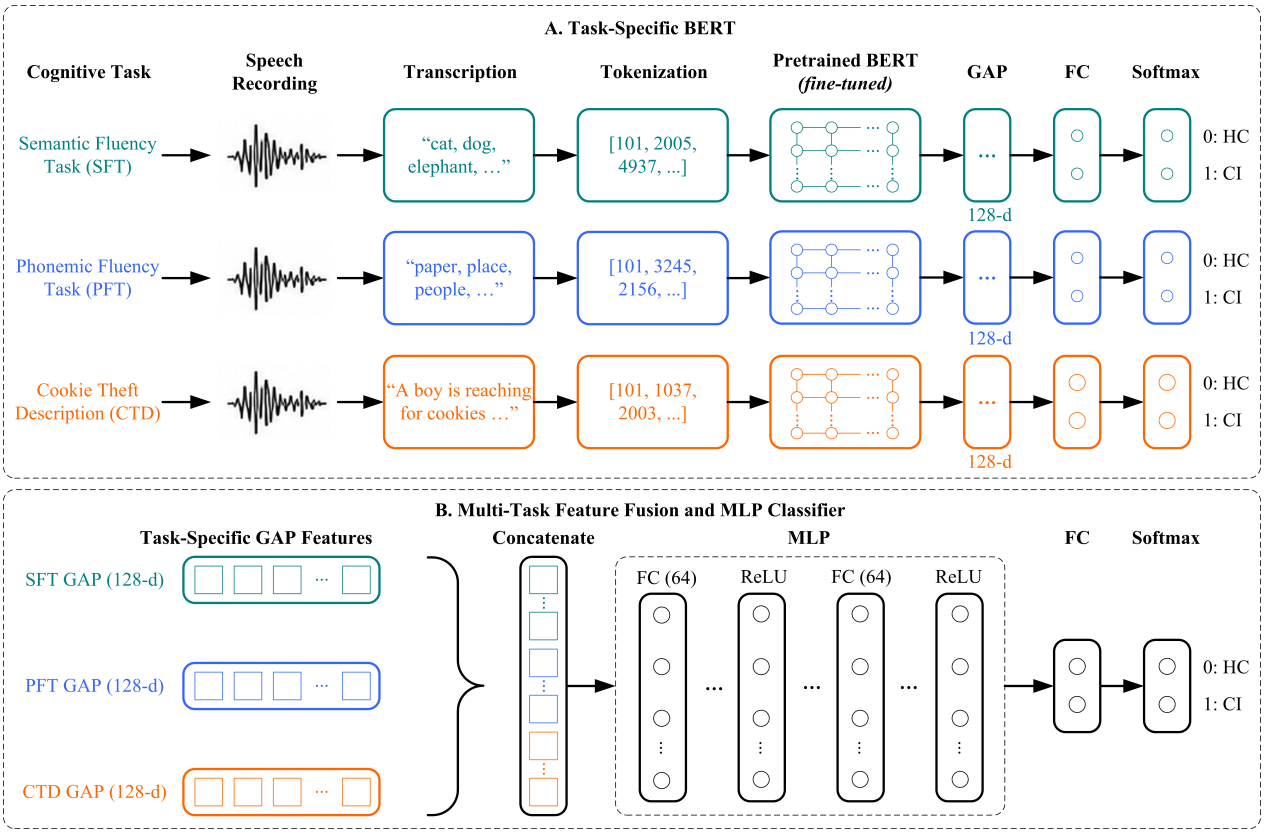

Fig. 1. Proposed framework for cognitive impairment detection from speech transcripts. Task-specific BERT encoders are independently fine-tuned using transcripts from the Semantic Fluency Task (SFT), Phonemic Fluency Task (PFT), and Cookie Theft Description (CTD) task. Global Average Pooling (GAP) is applied to obtain 128-dimensional embeddings from each task, which are concatenated and classified using a multilayer perceptron (MLP) to discriminate cognitively impaired (CI) participants from healthy controls (HC).

$$z = \frac{1}{T}\sum_{t=1}^{T} h_t, \tag{1}$$

where $z \in \mathbb{R}^{128}$ represents a compact semantic representation of the transcript. Compared with using individual token embeddings, the GAP operation aggregates contextual information across the entire sequence while maintaining a fixed-dimensional feature vector suitable for downstream classification.

The resulting 128-dimensional embedding is connected to a fully connected classification layer followed by a softmax activation function [14]. Separate BERT models are fine-tuned for the SFT, PFT, and CTD tasks using the binary CI versus HC classification objective.

C. Feature Fusion and Classification

The three cognitive assessment tasks included in PROCESS-2 probe different aspects of cognitive and linguistic functioning. Specifically, SFT evaluates semantic memory and lexical retrieval, the PFT assesses phonological retrieval and executive control, and the CTD task captures spontaneous discourse production. Consequently, each task may provide complementary information for identifying cognitive impairment. To exploit this complementary information, task-specific embeddings learned by the individual BERT encoders are combined through feature-level fusion.

For each participant, the fine-tuned SFT, PFT, and CTD encoders generate transcript-level embeddings denoted by z_{SFT}, z_{PFT}, and z_{CTD}, respectively, where $z_{\mathrm{SFT}}, z_{\mathrm{PFT}}, z_{\mathrm{CTD}} \in \mathbb{R}^{128}$. The three embeddings are concatenated to form a unified feature representation

$$z_f = [z_{\mathrm{SFT}}; z_{\mathrm{PFT}}; z_{\mathrm{CTD}}], \tag{2}$$

where $z_f \in \mathbb{R}^{384}$ denotes the fused feature vector. This representation integrates semantic information extracted from all three cognitive tasks and serves as the input to the final classifier.

Classification is performed using a MLP consisting of two hidden fully connected layers, each containing 64 neurons and followed by a rectified linear unit (ReLU) activation function [15]. The output layer contains two neurons corresponding to the HC and CI classes. A softmax activation function is applied to produce posterior class probabilities.

D. Baseline Methods

To provide a benchmark for evaluating the proposed transformer-based framework, several traditional machine learning classifiers were implemented using lexical features derived from the speech transcripts. The baseline systems

employed a bag-of-words representation, which has been widely used in text classification tasks and serves as a strong conventional approach for linguistic analysis.

For each transcript, textual preprocessing was first performed by tokenizing the document and removing stop words. A bag-of-words model was subsequently constructed using the training set transcripts, where each feature corresponded to the occurrence count of a word within the vocabulary. To reduce vocabulary sparsity and the influence of infrequent terms, words with a occurrence count of two or fewer were removed from the bag-of-words vocabulary. The resulting vocabulary was then used to encode the validation and test sets, ensuring that feature extraction was performed consistently across all data partitions.

Feature selection was applied prior to classifier training. Specifically, a two-sample t-test was performed on the training data to identify features exhibiting statistically significant differences between the CI and HC groups. Features with a p-value less than 0.05 were retained for subsequent classification.

Five conventional machine learning classifiers were evaluated using the selected lexical features: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). These models provide a diverse set of linear and non-linear classification approaches and serve as reference methods for assessing the effectiveness of the proposed BERT-based feature learning and fusion framework.

E. Experimental Setup

The experiments were conducted using the predefined training and testing partitions of the PROCESS-2 dataset. The training partition consisted of 320 participants, while the testing partition contained 80 participants. To facilitate model selection and hyperparameter tuning, 20% of the training data was randomly reserved as a validation set. The resulting validation subset was used for monitoring model performance during training, whereas the test set was used exclusively for final performance evaluation.

For the proposed transformer-based framework, separate BERT models were fine-tuned for the SFT, PFT, and CTD tasks. Transcript sequences were tokenized using the pretrained BERT tokenizer. Model training was performed using the Adam optimizer [16] with an initial learning rate of 1×10^{-4}, a mini-batch size of 8, and a maximum of 10 training epochs. A piecewise learning rate schedule was employed, with the learning rate reduced by a factor of 0.05 every five epochs. Network parameters were optimized using categorical cross-entropy loss [14].

For the feature fusion stage, the 128-dimensional embeddings extracted from the three task-specific encoders were concatenated to form a 384-dimensional feature vector. The resulting MLP classifier was trained using stochastic gradient descent with momentum (SGDM) [17], an initial learning rate of 1×10^{-4}, a mini-batch size of 8, and a maximum of 10 epochs. The learning rate was reduced by a factor of 0.5 every five epochs to facilitate stable convergence.

Performance was evaluated on the binary classification task of distinguishing CI from HC participants. Classification performance was assessed using standard evaluation metrics including recall, precision, and F1-score. All these metrics are reported in percentage (%).

III. EXPERIMENTAL RESULTS

All results reported in this section correspond to performance on the held-out PROCESS-2 test set.

A. Performance on Individual Cognitive Tasks

Table II presents the classification performance obtained using traditional machine learning classifiers and the proposed BERT-based approach for each cognitive speech task. Across all three tasks, the BERT models achieved the highest F1-scores, demonstrating the effectiveness of contextual language representations for cognitive impairment detection.

For SFT, the proposed BERT model achieved an F1-score of 83.95%, outperforming the best traditional machine learning baseline, namely RF, which obtained an F1-score of 78.48%. Similar improvements were observed for PFT, where BERT achieved an F1-score of 82.93% compared with 78.48% obtained by RF and 77.65% obtained by SVM. These results suggest that transformer-based contextual representations are better able to capture semantic and lexical characteristics associated with cognitive impairment than conventional bag-of-words features.

Among the individual tasks, CTD task yielded the strongest performance. The BERT model achieved an F1-score of 84.71%, exceeding all traditional machine learning approaches. The superior performance of CTD may be attributed to the richer linguistic content available in connected speech, which provides more opportunities for the model to capture cognitive-linguistic deficits than constrained verbal fluency tasks. In contrast, SFT and PFT responses are generally shorter and more structured, resulting in less contextual information for representation learning.

B. Effect of Feature Fusion

Table III summarizes the performance of the task-specific BERT models and the proposed feature fusion framework. While each cognitive task independently provided useful information for identifying cognitive impairment, combining task-specific embeddings resulted in further performance improvements.

The best individual task performance was obtained using the CTD-based BERT model, which achieved an F1-score of 84.71%. By concatenating embeddings extracted from the SFT, PFT, and CTD tasks and performing classification using an MLP, the proposed fusion framework achieved an F1-score of 87.06%. This corresponds to an absolute improvement of 2.35% over the strongest single-task model.

The fusion model also achieved the highest recall of 92.50%, indicating improved sensitivity for detecting cognitively impaired participants. These findings suggest that the three cognitive tasks capture complementary aspects of cognitive and linguistic functioning and that integrating

TABLE II

TEST-SET PERFORMANCE (%) OF TRADITIONAL MACHINE LEARNING CLASSIFIERS AND BERT-BASED MODELS ACROSS THE SEMANTIC FLUENCY TASK (SFT), PHONEMIC FLUENCY TASK (PFT), AND COOKIE THEFT DESCRIPTION (CTD) TASK

Method	SFT			PFT			CTD		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
LR	75.00	63.83	68.97	72.50	65.91	69.05	82.50	84.62	83.54
LDA	65.00	70.27	67.53	77.50	68.89	72.94	72.50	64.44	68.24
DT	62.50	73.53	67.57	72.50	76.32	74.36	75.00	68.18	71.43
RF	77.50	79.49	78.48	77.50	79.49	78.48	82.50	84.62	83.54
SVM	75.00	75.00	75.00	82.50	73.33	77.65	82.50	82.50	82.50
BERT	85.00	82.93	83.95	85.00	80.95	82.93	90.00	80.00	84.71

TABLE III

TEST-SET PERFORMANCE (%) OF TASK-SPECIFIC BERT MODELS AND THE PROPOSED FEATURE FUSION FRAMEWORK FOR BINARY COGNITIVE IMPAIRMENT CLASSIFICATION

Model	Recall	Precision	F1-Score
SFT-BERT (Task-specific)	85.00	82.93	83.95
PFT-BERT (Task-specific)	85.00	80.95	82.93
CTD-BERT (Task-specific)	90.00	80.00	84.71
Proposed Fusion (MLP)	92.50	82.22	87.06

information across tasks provides a more comprehensive representation of participant status than any individual task alone.

IV. DISCUSSION AND CONCLUSION

The results demonstrate two important findings. First, transformer-based language models outperformed traditional machine learning approaches based on bag-of-words features across all cognitive tasks. This highlights the value of contextual semantic representations for modeling subtle linguistic changes associated with cognitive impairment.

Second, the proposed feature fusion strategy improved classification performance by integrating information from multiple cognitive assessments. The observed performance gain supports the hypothesis that SFT, PFT, and CTD evaluate different cognitive processes and therefore provide complementary information for automatic cognitive impairment detection. While CTD emerged as the strongest individual task, the fusion framework achieved the best overall performance, indicating that useful discriminative information remains present in the verbal fluency tasks.

Overall, the findings suggest that combining task-specific transformer representations from multiple cognitive speech assessments is a promising approach for speech-based cognitive screening and may provide a more robust characterization of cognitive impairment than relying on a single assessment task.

REFERENCES

[1] J. N. Vega and P. A. Newhouse, "Mild cognitive impairment: Diagnosis, longitudinal course, and emerging treatments," *Current Psychiatry Reports*, vol. 16, 2014, Art. no. 490.

[2] S. A. Eshkoor, T. A. Hamid, C. Y. Mun, and C. K. Ng, "Mild cognitive impairment: Diagnosis, longitudinal course, and emerging treatments," *Clinical Interventions in Aging*, vol. 10, pp. 687–693, 2015.

[3] M. N. Sabbagh, M. Boada, S. Borson *et al.*, "Rationale for early diagnosis of mild cognitive impairment (MCI) supported by emerging digital technologies," *The Journal of Prevention of Alzheimer's Disease*, vol. 7, no. 3, pp. 158–164, 2020.

[4] N. R. Fowler, K. A. Partrick, J. Taylor *et al.*, "Implementing early detection of cognitive impairment in primary care to improve care for older adults," *Journal of Internal Medicine*, vol. 298, no. 1, pp. 31–45, 2025.

[5] S. Gauthier, P. Rosa-Neto, J. A. Morais, and C. Webster, *World Alzheimer Report 2021: Journey Through the Diagnosis of Dementia*. London, England: Alzheimer's Disease International, 2021.

[6] K. C. Fraser, J. A. Meltzer, F. Rudzicz, and P. Garrard, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[7] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 2162–2166.

[8] R. V. Sharan, H. Xiong, and Y.-J. Tseng, "Fusion of traditional and large language models for linguistic analysis in Alzheimer's disease detection," in *Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1–5.

[9] M. Pahar, C. H. Illingworth, B. Mirheidari *et al.*, "PROCESS-2: A benchmark speech corpus for early cognitive impairment detection," *arXiv:2605.14888*, 2026.

[10] R. M. Vaughan, R. F. Coen, R. Kenny, and B. A. Lawlor, "Semantic and phonemic verbal fluency discrepancy in mild cognitive impairment: Potential predictor of progression to Alzheimer's disease," *Journal of the American Geriatrics Society*, vol. 66, no. 4, pp. 755–759, 2018.

[11] V. A. G. Steiner, L. L. Mansur, S. M. D. Brucki, and R. Nitrini, "Phonemic verbal fluency and age: A preliminary study," *Dementia & Neuropsychologia*, vol. 2, no. 4, pp. 328–332, 2008.

[12] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task," *Neurological Sciences*, vol. 26, pp. 243–254, 2005.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015.

[17] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1139–1147.