IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 January 2026, accepted 26 January 2026, date of publication 6 February 2026, date of current version 12 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3662045

RESEARCH ARTICLE

The DigiDiaDem Speech-Cognitive Dataset: Initial Experiments on Detecting Cognitive Impairments From Speech

LUBOŠ ŠMÍDL[1], (Member, IEEE), FILIP POLÁK[2], LUCIE ZAJÍCOVÁ[2], TOMÁŠ LEBEDA[2], JAN ŠVEC[1], JAN TUPÝ[2], MARTIN BULÍN[2], AND ALEŠ BARTOŠ[3,4]

[1]Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia in Pilsen, Pilsen 30100, Czech Republic
[2]NTIS Research Center, Faculty of Applied Sciences, University of West Bohemia in Pilsen, Pilsen 30100, Czech Republic
[3]Department of Neurology, Third Faculty of Medicine, Charles University, Prague 10000, Czech Republic
[4]Department of Neurology, University Hospital Kralovske Vinohrady, Prague 10000, Czech Republic

Corresponding author: Luboš Šmídl (smidl@kky.zcu.cz)

This research was supported by the Technology Agency of the Czech Republic, SIGMA project No. TQ01000332. Data collection was supported by Neuroscience COOPERATIO [207038] from Charles University, and by Development of research organizations from the Ministry of Health [Kralovske Vinohrady University Hospital, 00064173]. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

ABSTRACT Dementia is a growing global health challenge, making early detection of cognitive decline critically important. Early detection of dementia and other cognitive impairment is essential for timely intervention and better care planning. However, existing datasets for training automated screening tools are limited, especially for underrepresented languages such as Czech. In this study, we present a new dataset and a novel application named the DigiDiaDem (Digital Diagnostics of Dementia) designed for automated dementia screening through multimodal cognitive assessment. The application integrates user-friendly digital cognitive tasks with machine learning algorithms to evaluate linguistic and cognitive performance in real time. Using this system, we collected and curated a comprehensive dataset of speech and cognitive data from Czech-speaking participants. The dataset comprises 371 individuals, including cognitively normal individuals and patients with mild cognitive impairment and mild dementia. It includes socio-demographic data, results of cognitive and speech tests, functional assessment questionnaires, data collected through the DigiDiaDem application, and automatic speech recognition (ASR) transcripts of spoken responses. Raw audio recordings are not included. Instead, the dataset provides manually engineered linguistic and acoustic features. We describe the data collection process and outline the cognitive tasks used to collect the dataset. Our experiments demonstrate that speech features derived from cognitively demanding tasks, such as verbal fluency and memory recall, can effectively distinguish healthy participants from those with cognitive impairment. Models trained on the dataset achieved up to 95% accuracy when combining speech features with demographic information. Preliminary experiments demonstrate the feasibility of using the collected data for dementia detection. These findings confirm that speech-based digital assessment can complement traditional clinical evaluation. The proposed dataset and application offer a substantial resource for the research community by establishing a solid baseline for machine-learning-based approaches to dementia screening from speech-based interaction.

INDEX TERMS Cognitive impairment, cognitive tests, dementia screening, dementia diagnostics, Alzheimer's disease, DigiDiaDem, classification features, machine learning algorithms, automatic speech recognition.

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

I. INTRODUCTION

The global healthcare landscape is at a critical inflexion point. Demographic shifts, such as aging populations and the rising

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

prevalence of chronic diseases, place unprecedented pressure on healthcare systems around the world. Conditions such as cardiovascular disease, diabetes, and neurodegenerative disorders, including dementia, require continuous long-term management. Traditional episodic models of care are not designed to support this type of management [1].

The World Health Organization projects that by 2050, the global population aged 60 years and older will double to 2.1 billion, increasing the demand for accessible and sustainable healthcare solutions [2]. According to estimates, the global population of patients is expected to triple by 2050 [3]. It is essential for society to develop innovative solutions to enhance the living conditions of the elderly population, particularly those who live alone [4]. The concurrent explosion in digital technology and computational power offers a powerful set of tools to reimagine how, when, and where healthcare is delivered.

Dementia belongs to the conditions most profoundly affected by this demographic transformation. It is a progressive neurodegenerative syndrome that impairs memory, cognition, and functional ability. Currently, more than 55 million people live with dementia worldwide, with nearly 10 million new cases reported each year [5]. Despite its growing prevalence, dementia often remains underdiagnosed or misdiagnosed, particularly in low-resource settings where access to specialized clinical assessments is limited [6]. Conventional diagnostic methods, which are mainly based on in-person cognitive testing and neuroimaging, are not only resource-intensive, but also impractical for large-scale or continuous monitoring. Consequently, there is a critical need for scalable, accessible, and data-driven approaches to support early detection and longitudinal tracking of cognitive decline. Digital health technologies—encompassing mobile applications, wearable sensors, and artificial intelligence (AI) — offer a promising avenue to bridge this diagnostic gap by enabling remote, unobtrusive, and cost-effective assessment of cognitive function. Recent research has increasingly focused on quantifiable changes in speech and language as powerful, early indicators of cognitive decline. The subtle shifts in semantic content, syntactic complexity, and fluency that accompany neurodegeneration can serve as potent digital biomarkers.

A. RELATED WORK

Traditional diagnostic tools for Alzheimer's disease (AD), including brain positron emission tomography and cerebrospinal fluid examination, remain the gold standard for clinical diagnosis but are costly, invasive, and impractical for routine or large-scale screening. The logistical and financial burden of these biomarker assessments underscores the need for alternative approaches that enable earlier, more accessible detection of cognitive decline. In recent years, there has been a paradigm shift toward digital and remote cognitive assessments that leverage subtle linguistic and cognitive changes as early, non-invasive indicators of neurodegeneration. These approaches promise to extend monitoring beyond specialized clinics, offering scalable, low-cost tools suitable for widespread deployment.

A growing body of research highlights the diagnostic potential of speech and language as rich sources of digital biomarkers. Subtle alterations in semantic content, syntactic complexity, and fluency can precede overt clinical symptoms, making connected speech analysis a particularly sensitive measure of cognitive decline. Picture description tasks have become a central method for eliciting structured yet naturalistic language samples, offering a standardized means to examine these linguistic features. Picture description tasks rely more on the individual's semantic knowledge and lexical retrieval than on episodic memory, thereby isolating the cognitive-linguistic domains most affected by AD [7]. The widely used ''Cookie Theft'' picture from the Boston Diagnostic Aphasia Examination provides a consistent context for cross-sectional and longitudinal comparisons [8]. Studies have shown that individuals with AD produce fewer ''information units,'' leading to globally less informative descriptions, and often demonstrate deficits in expressing a picture's central themes [9], [10]. Moreover, AD-related language decline manifests through simplified syntax and a higher frequency of word-finding difficulties, particularly evident in written descriptions where grammatical errors and semantic intrusions are more prevalent [9]. The literature review focused on clinically validated cognitive assessment tools, speech-based screening approaches, and digital datasets relevant to dementia detection.

The integration of artificial intelligence and machine learning (ML) has accelerated progress in this field by enabling the automated and consistent analysis of linguistic and acoustic features. These computational approaches allow the detection of patterns more nuanced or complex for consistent human evaluation. Deep learning models have been successfully applied to speech from narrative recall tasks, distinguishing individuals with mild cognitive impairment (MCI) or mild AD from cognitively unimpaired individuals with high accuracy [11]. Similarly, conversational data analyzed using traditional ML algorithms have achieved accuracies exceeding 80% in classifying cognitive impairment [12], while ensemble learning approaches such as AdaBoost have reached accuracies up to 92.5% in remote assessments [13]. These results underscore the robustness of speech-based machine learning as a diagnostic modality.

Building upon these advances, researchers have increasingly focused on deploying such assessments via remote and mobile platforms. The move toward digital cognitive testing addresses longstanding barriers to access, particularly in rural or underserved regions where specialized neurological services are limited [14]. Remote assessments delivered through smartphones and tablets have demonstrated strong equivalence with in-clinic testing, as evidenced by high correlations between face-to-face and video-conference-administered word recall and naming tasks [14]. Feasibility studies confirm that participants can complete these assessments independently. Speech tasks performed via smartphones were

associated with 86% adherence [11]. Another study has shown that brief, tablet-based speech assessments effectively distinguish cognitively normal individuals from those with MCI or AD [15]. Complementary developments in secure cloud infrastructures further enable scalable data processing and storage for these remote systems [13]. Beyond AD, similar AI-driven telemedicine models have been applied to aphasia diagnosis and treatment [16], illustrating the versatility and clinical validity of remote, voice-based assessment.

Collectively, this literature demonstrates two converging trends: the maturation of linguistic biomarkers as indicators of cognitive decline and the growing feasibility of remote, AI-driven assessment. The present work builds upon these foundations by integrating speech-based analysis within a mobile platform for dementia detection, combining the validated sensitivity of connected speech measures with the accessibility and scalability of digital health technology.

Several well-established instruments, such as the Cohen-Mansfield Agitation Inventory (CMAI) and the Pittsburgh Agitation Scale (PAS), have been developed to evaluate behavioral symptoms of dementia, particularly agitation and aggression in moderate to advanced stages of the disease [17], [18]. While these tools play a vital role in clinical monitoring and care planning, they are not designed for early stage cognitive screening. Our work complements these approaches by focusing on the detection of subtle speech and task related impairments that characterize the prodromal phases of Alzheimer's disease and related dementias. As such, DigiDiaDem is not intended to replace behavioral scales but rather to support earlier identification using accessible and interpretable speech-based measures.

B. PREVIOUS WORK

Prior to commencing the main development, preliminary experiments were conducted to assess whether automatic speech recognition (ASR) could achieve reasonable accuracy in recognizing patient recordings in Czech [19]. This step was crucial since the overarching goal was to enable patient classification without requiring a specialist or doctor to manually analyze the recordings or data. The results were promising, thus enabling us to start the whole DigiDiaDem (Digital Diagnostics of Dementia) project. These experiments are further explored in Section III-D.

In our previous studies, we demonstrated the feasibility of classification based on semantically or acoustically derived features. In the first work, we introduced a system that uses expert-annotated semantic description of images serving as ground truth, with a custom semantic parser to compare natural language descriptions with this reference annotation [20]. This comparison yielded an interpretable semantic loss vector, which was subsequently used for classification. In the second study, we focused on the repetition of pseudowords and utilized acoustic features and similarity measures to derive an analogous loss vector for classification [21]. Building on these preliminary results, the present paper aims to provide a comprehensive evaluation of this approach across all tasks.

C. PAPER ORGANIZATION

In the following sections, we introduce the DigiDiaDem application (Section II), describe the collected dataset (Section III), present the baseline experimental results (Section IV), and conclude with the potential impact of the dataset (Section VI).

This paper focuses on presenting the open, publicly-available DigiDiaDem Speech-Cognitive Dataset [22], including the methodology of its collection, the structure of the cognitive and speech tasks, and the baseline results for automated dementia detection in Czech-speaking populations. The goal is to provide a reproducible benchmark for future studies. While we demonstrate initial classification performance using standard machine learning models, the paper does not aim to exhaustively explore feature selection, model optimization, or clinical deployment, which are left for future work.

II. OVERVIEW OF THE DigiDiaDem APPLICATION

We developed the DigiDiaDem, a digital application designed to facilitate accessible, data-driven screening and monitoring of dementia. The platform evaluates speech and memory tasks, such as memory recall, language comprehension, and executive function assessments, using an intuitive web-based interface. Responses are collected as audio recordings. These audio recordings are then transcribed into text, from which performance-based data derived from task results are extracted. This provides a rich foundation for computational analysis. Using these data, the DigiDiaDem employs machine learning algorithms to identify subtle patterns associated with cognitive decline, offering the potential for early detection even before overt clinical symptoms emerge. Beyond serving as a screening tool, the application acts as a research framework, supporting the collection of a structured, ethically sourced dataset that advances dementia-focused AI models. This dual focus on clinical utility and data generation positions the DigiDiaDem as a scalable and privacy-conscious contribution to the growing field of digital cognitive health.

The DigiDiaDem application is a spoken dialog system designed to facilitate interactive communication with users. A central aim of the DigiDiaDem was the transformation of traditional in-person cognition pen-and-paper assessments into a fully digital format. The rationale for this transition includes:

- **Increased accessibility:** Computers and mobile devices are widely available, even among older populations, often with the help of family members or caregivers.
- **Scalability and remote administration:** A digital platform enables remote testing, thereby removing geographical and logistical barriers.

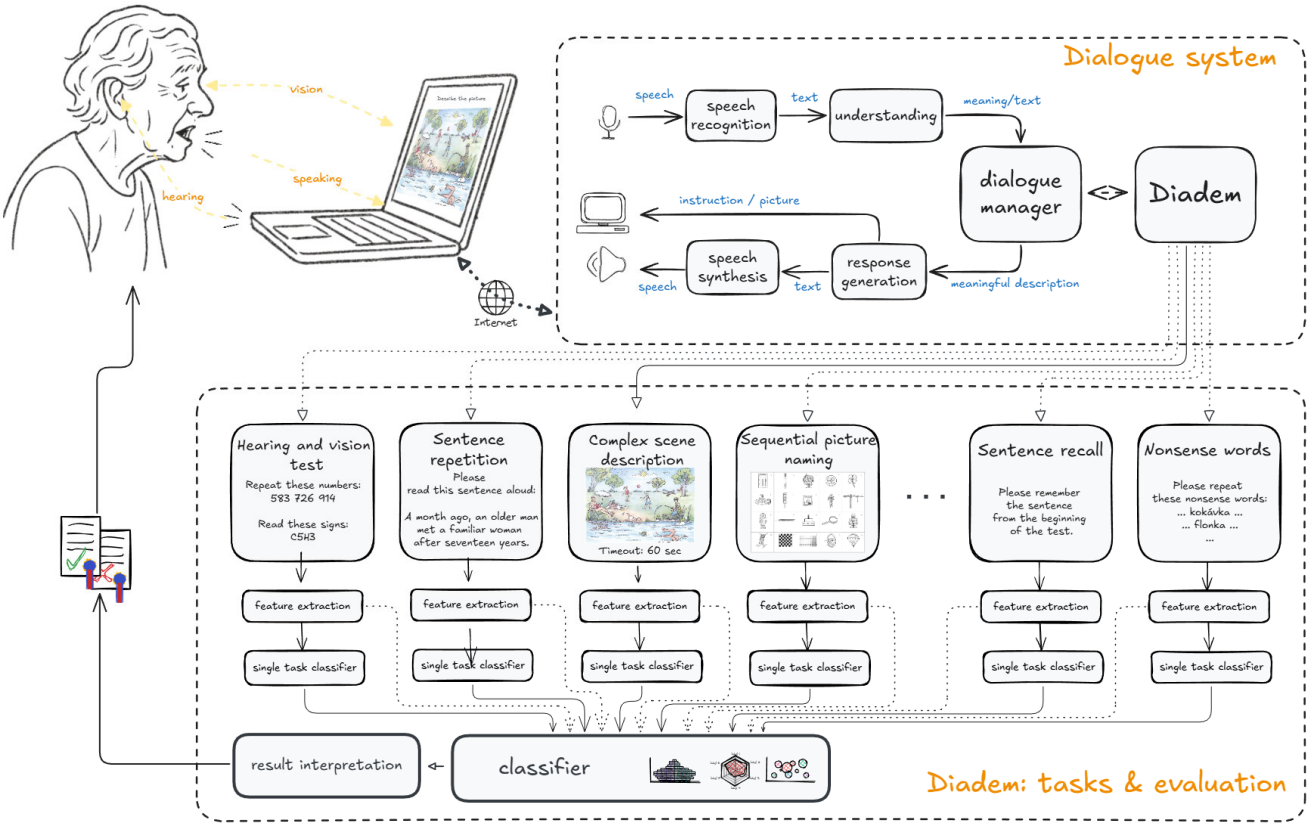


FIGURE 1. DigiDiaDem application overview diagram.

- **Utilisation of assessment results:** Traditional pen-and-paper cognitive assessments are a powerful tool to detect cognitive impairment. By combining the results of these assessments with the speech analysis, we can obtain even more robust results.

From the outset, the development team focused on accessibility, usability, and clinical validity. A key design decision was to leverage speech as the primary means of interaction between the user and the system. This choice was motivated by several considerations:

- **Accessibility for older adults:** Individuals with mild cognitive impairment (MCI) or mild dementia often retain the ability to speak, while fine motor tasks, such as typing or using a stylus, may present significant challenges.
- **Minimising user burden:** A speech-driven interface allows for intuitive interactions, reducing the cognitive and physical demands.
- **Naturalistic data collection:** Speech provides a rich source of behavioral and linguistic markers, allowing the system to analyze subtle changes in prosody, fluency, and vocabulary that may indicate early cognitive decline.

The objective of users is to access the application using a computer, complete the examination, and subsequently receive feedback on their performance. The feedback is generated by the machine learning pipeline based on speech analysis and the results of individual tasks. The classification could be multi-class (reflecting the target classes from the training dataset) or binary (with/without cognitive impairment). If certain markers of cognitive decline are detected, users are advised to contact a specialist. If no markers are detected, users are advised to repeat the examination after a specified interval.

A. COGNITIVE TASKS OVERVIEW

The DigiDiaDem test battery comprises 10 tasks designed to assess various sensory and cognitive abilities. These tasks include:

1) Hearing examination - repetition of numbers, which serves as a hearing test.
2) Vision examination - reading characters to evaluate eyesight.
3) Sentence repetition - repeating and encoding a sentence.
4) Complex scene description - describing a hand-drawn picture.
5) Complex scene object recall - recalling as many objects depicted in the picture as possible.
6) Sequential picture naming - describing twenty hand-drawn black-and-white pictures.

7) Picture recall - recalling as many of the twenty pictures as possible.
8) Animal verbal fluency - naming as many animals as possible.
9) Sentence recall - recalling the sentence from the task at the beginning.
10) Nonsense words repetition - repeating nonsense words.

The system was designed to be a comprehensive data collection tool, which means that each task was chosen to measure specific types of cognitive goals. The tasks are also created as modular components of spoken dialog systems, allowing them to be easily swapped, removed, or replaced with other tasks (even those designed to test a different type of impairment, such as Parkinson's disease, for example).

In Figure 1, individual tasks from the test battery are visualized as square boxes. The figure also includes the classifier and result interpretation layers, which utilize the extracted task features to determine whether the user has cognitive impairment and present the results to the user.

1) TASK 01: HEARING EXAMINATION

The first test administered to users is the number-repeating task. In this task, users are asked to repeat three-digit numbers spoken aloud by a text-to-speech (TTS) system. The specific numbers used are 583, 726, and 914.

This test is designed primarily to serve as a hearing and comprehension check. It allows verification that the volume level is appropriate for the user and that they can clearly understand the TTS voice.

2) TASK 02: VISION EXAMINATION

In this task, users are asked to read aloud four characters displayed on the screen: `C 5 H 3`, pronounced in Czech as *"cé pět há tři"* ([tseː pjɛt haː tr̝i]). These characters are presented in decreasing font size to test the user's eyesight and ensure that they will not experience difficulty reading the text throughout the examination.

Primarily, this task serves as a simple vision screening to ensure that the user can physically see and read the text required for subsequent tests. However, beyond its function as an eyesight check, it also engages several fundamental cognitive processes that are crucial for successful completion:

- **Character Recognition:** The ability to correctly identify a mix of letters and numbers (`C`, `5`, `H`, `3`) tests basic visual processing and recognition. The brain must not only perceive the shapes, but also correctly map them to their symbolic meaning as known characters. Failure to identify a clearly visible character could suggest issues with visual gnosis (the ability to recognize objects) rather than just eyesight.
- **Naming and Verbal Production:** Reading the characters aloud relies on the ability to name them, a core function of expressive language. This simple act requires retrieving the name of the symbol from memory and articulating it. Significant hesitation or inability to name a recognized character could be an early sign of anomia (word-finding difficulty), a common symptom in various forms of dementia [23].

3) TASK 03: SENTENCE REPETITION

In this task, the user is asked to repeat a sentence played as an audio recording. The sentence used is: *Před měsÍcem starší muž potkal známou ženu po sedmnácti letech.* (English translation: *A month ago, an older man met a familiar woman after seventeen years.*).

The order of words in this sentence is intentionally unconventional to increase cognitive load, encouraging users to engage more actively in working memory. Furthermore, the sentence must be remembered for a delayed recall task after approximately ten minutes (seven tasks later), adding an element of memory storage assessment.

This task imposes greater cognitive demands than typical memory tests, most of which assess recall of individual words rather than entire sentences [24], [25]. The number of such stimuli is relatively low, ranging from two to seven words or items. It usually takes 10 to 30 minutes to yield memory scores for the various cognitive tests [26], [27].

This task builds on the clinically validated ALBA (Amnesia Light and Brief Assessment) protocol [28], which uses sentence repetition to assess verbal memory and linguistic integrity in patients with suspected cognitive decline. Unlike common memory tests that use isolated words, the ALBA approach employs sentence-level stimuli to increase cognitive demand through syntactic structure and semantic integration. In our adaptation, we extended the sentence length from six to ten words to further challenge memory span and expressive language.

- **Verbal Working Memory:** The primary function tested is the capacity of the phonological loop, the component of working memory responsible for temporarily storing and rehearsing auditory information. Holding a 10-word sentence, with its specific vocabulary and grammatical structure, places a much greater load on this system than repeating three digits. Errors such as word omissions, substitutions, or sequencing mistakes can indicate a limited working memory capacity.
- **Language Processing:** Beyond just memory, this task requires intact language skills. The user must correctly process the sentence's syntax (grammatical structure) and semantics (meaning) to effectively "chunk" the information, which aids in recall. The ability to articulate the sentence correctly also assesses expressive language functions. The difficulty in this task could signal early signs of aphasia, a language disorder that can accompany dementia.
- **Attention and Concentration:** To encode the entire sentence accurately, the user must maintain focused attention for its full duration. Unlike a short digit string, a longer sentence is more susceptible to interference

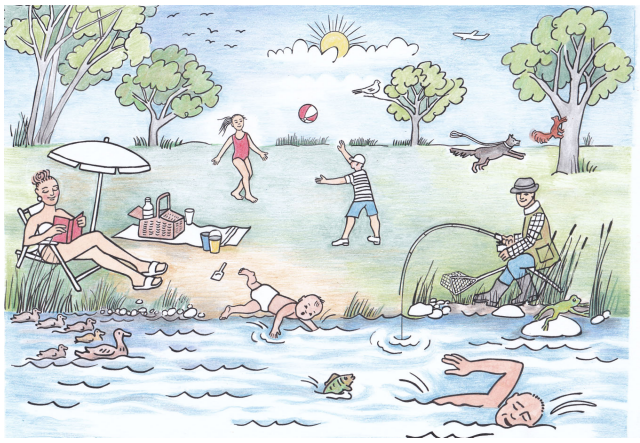


FIGURE 2. Colored vivid shore picture.

from attentional lapses, which could manifest as forgetting the beginning or end of the sentence.

4) TASK 04: COMPLEX SCENE DESCRIPTION

The user is asked to describe a hand-drawn colored scene displayed in Figure 2 in as much detail as possible in 60 seconds.

The design of the picture reflects considerations related to cognitive disorders, such as dementia. Cognitive impairment not only affects the memory of individual words, but also significantly impairs the patient's ability to maintain a coherent flow of thought and describe interconnected events within the scene.

The scene is deliberately constructed with three layers of interest:

1) the water of the lake in the foreground,
2) the grassy shoreline in the middleground,
3) the sky in the background.

Each layer contains multiple events, ensuring that the subject has continuous material to describe. The division of layers and the distribution of events were carefully designed to facilitate the analysis of the patient's narrative flow. Specifically, we investigate:

- Where do users begin and end their description of the picture?
- Do they navigate the scene horizontally or vertically?
- Do they notice and describe all the events?
- Do they mention the danger situation (i.e., child falling into water)?
- Do they use isolated words or form connected descriptions?
- Do they perceive the overall scene (e.g., "*a sunny day at the river*")?
- Do they remember specific and accurate words (e.g., "*squirrel*")?
- Does the described story follow a coherent storyline, or do they merely repeat a few isolated words describing the most obvious activities?

This task is a comprehensive assessment that evaluates the integration of high-level cognitive functions, moving far beyond simple object naming. The requirement of generating a detailed narrative from a complex visual scene is a classic method in neuropsychology, famously exemplified by the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination [29]. The deliberate three-layered structure of the scene is designed to probe several specific cognitive domains:

- **Executive Function and Organizational Strategy:** A primary goal is to observe the user's strategic approach. A cognitively intact individual might describe the scene systematically (e.g., by describing the foreground first, then the middleground, and finally the background, or by moving from left to right). In contrast, individuals with executive dysfunction, often seen in frontal lobe dementia, can produce a disorganized, fragmented, or chaotic description, jumping between unrelated elements without a clear plan. The ability to shift attention between layers while maintaining a coherent narrative evaluates cognitive flexibility.
- **Language Production and Narrative Coherence:** This directly assesses expressive language skills. The analysis goes beyond simple word-finding (lexical retrieval) to evaluate syntax (grammatical correctness), semantic richness (variety of vocabulary), and, most importantly, discourse coherence. The task reveals whether the user can simply list items (*"I see a tree, a dog, a man"*) or if they can weave these elements into a meaningful narrative (*"A man is catching fish at the lake, and there are tall trees on the shore behind him"*). A breakdown in this narrative ability may serve as a sensitive marker of cognitive decline.
- **Visuospatial Interpretation:** Before any description can be generated, the user must correctly perceive and interpret the visual information. This includes identifying individual objects, understanding their relationships to one another, and appreciating the overall context of the scene. The multi-layered design also subtly evaluates the ability to process depth cues and perceive a three-dimensional space from a two-dimensional image. Errors in identifying objects or describing their spatial locations can indicate deficits in this domain [30].
- **Sustained Attention:** The 60-second time limit requires the user to maintain focus on the task. The quantity and quality of the description can provide insight into the person's ability to sustain attention without becoming distracted or prematurely ceasing the task.

5) TASK 05: COMPLEX SCENE OBJECT RECALL

This task follows directly after the description of the picture of the shore scene. The user is asked to recall and list as many items as they can recall from the scene they have just seen. The time limit for this task is 30 seconds.

The primary goal of this task is to evaluate the user's short-term memory by measuring how many and which elements of the image they can recall. Thanks to the deliberately

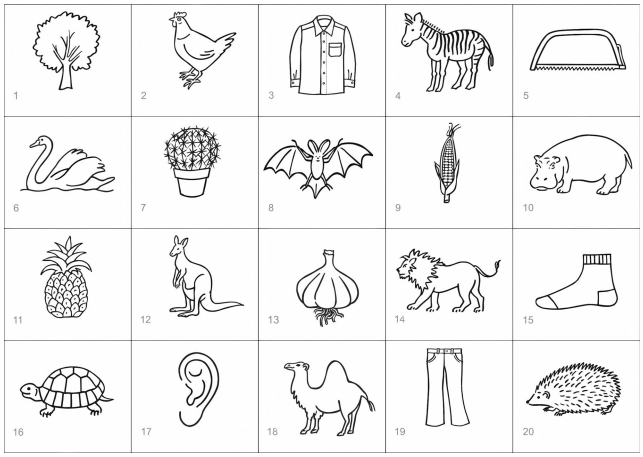


FIGURE 3. Sequential naming of individual 20 black-and-white drawings.

layered structure of the scene (foreground, middleground, background), it is possible to determine which parts of the scene the user remembers mostly.

6) TASK 06: SEQUENTIAL PICTURE NAMING

In this task, the user is presented with a sequence of 20 individual black-and-white drawings, as shown in Figure 3. These pictures were carefully prepared and selected with important characteristics, which are described in detail in our previous study [31]. Each image is shown for a short time (5 sec), during which the user is expected to name the object or concept depicted using a single word.

This task is a classic test of confrontation naming, designed to directly assess lexical retrieval and identify signs of anomia (word-finding difficulty), a common symptom of dementia [29]. Semantic memory is quickly assessed by naming objects or pictures or by semantic questions as a standalone test [26], [32] or as part of a test battery [26], [27], [33], [34]. Written, oral, or typed picture naming tests can also be employed to evaluate language abilities [35]. Although it may seem simple, successfully naming an object requires a rapid and seamless integration of several cognitive processes.

- **Visual Recognition:** The user must first perceive the drawing and correctly identify the object, relying solely on form and contour, as the black-and-white format removes color cues.
- **Semantic Access:** After recognizing the object, the user must connect it to the corresponding concept stored in their semantic memory.
- **Lexical Retrieval:** The final, and often most challenging, step is to retrieve the specific verbal label or name for that concept from their mental lexicon and speak it aloud.

The timed presentation of each image adds a processing speed component, making the task sensitive not only to the accuracy of naming but also to the efficiency of these underlying cognitive systems.

7) TASK 07: PICTURE RECALL

The objective of the task is to assess the user's short-term memory by asking them to recall as many of the 20 previously shown black-and-white drawings as possible within a limited time frame of 30 seconds.

By comparing the user's recall performance with their responses from the prior naming task, we can gain insights into memory retention and word association:

- **Recall quantity:** How many pictures was the user able to recall in the 30-second window?
- **Recall sequence:** The order in which the pictures are recalled. For example, whether the user starts recalling from the beginning or the end of the sequence.
- **Recall accuracy:** Situations where the user misidentified a picture during the naming task but recalls the correct word here (e.g., previously naming picture 11 as "*rhinoceros*" but now correctly recalling "*hippopotamus*").

8) TASK 08: ANIMAL VERBAL FLUENCY (AFT)

This task is a standard verbal fluency test widely used in neuropsychological diagnostics. The user is asked to name as many animals as they can within a 30-second time frame. Unlike previous tasks that focused on immediate or short-term recall, this task targets long-term memory and lexical access [36].

The primary aim of this task is to evaluate the user's long-term memory, psychomotor speed, and verbal fluency, particularly within the semantic category of animals. Although the total number of animals named during the time limit is the most basic performance metric, the analysis extends to several additional qualitative factors:

- **Repetitions:** The number of repeated words is noted, as excessive repetition may indicate cognitive or linguistic impairment.
- **Category diversity:** Responses are analyzed by the diversity of animal groups mentioned, such as domestic animals, aquatic animals, insects, and birds. Broader category coverage may suggest a more flexible and intact lexical retrieval system.

This task is particularly useful for detecting early stages of Alzheimer's disease (AD), as it reflects degradation of semantic networks and executive control. It also helps differentiate AD from other dementia types, such as frontotemporal dementia (FTD), where semantic fluency is often preserved longer. Furthermore, AFT is considered culturally neutral, making it suitable for use across diverse populations.

9) TASK 09: SENTENCE RECALL

This task revisits the sentence that the user heard and repeated in Task 03. The goal is to assess delayed verbal recall, a cognitive function often impaired in individuals with memory-related disorders.

This delayed recall task is designed to evaluate three critical components of memory performance:

- **Encoding** – Whether the user successfully stored the sentence when it was first presented.
- **Retention** – Whether the user maintained the memory of the sentence across several intervening tasks.
- **Retrieval** – Whether the user can access and reproduce the stored information after a time delay.

10) TASK 10: NONSENSE WORDS REPETITION

This is the final task in the DigiDiaDem application. The user is asked to repeat a sequence of sixteen artificially constructed nonsense words. These words do not exist in any known lexicon and are created to be phonologically plausible but semantically meaningless. The purpose of this task is to test the user's ability to process and reproduce unfamiliar auditory sequences, a skill often compromised in individuals with dementia.

The nonsense words used in this task are as follows:

Word	IPA Transcription
čapiváček	[ˈtʃapivaːtʃɛk]
kokávka	[ˈkokaːfka]
flonka	[ˈflonka]
vekutalóna	[ˈvɛkutaloːna]
glonfrou	[ˈglonfrou]
bitonyp	[ˈbitonɪp]
nedimutovatel	[ˈnɛdimutovatɛl]
ziviprolátum	[ˈzɪviprolaːtum]
spekofidlo	[ˈspɛkofidlo]
kmufoň	[ˈkmufoɲ]
klamtaše	[ˈklamtaʃɛ]
bodchudýni	[ˈbotxudiːji]
lampešiška	[ˈlampɛʃiʃka]
trebice	[ˈtrɛbɪtsɛ]
cidla	[ˈtsidla]

This task primarily evaluates the user's ability to perceive, retain, and reproduce complex phonological information in the absence of semantic cues. Such tasks are cognitively demanding, relying heavily on auditory working memory, phonological processing, and speech-motor control. The difficulty of repeating nonsense words, particularly when phonetic complexity is moderate to high, has been identified as a strong diagnostic marker for cognitive decline and dementia [37].

III. DESCRIPTION OF THE DATASET

The dataset has been collected in the past three years through the DigiDiaDem spoken dialog system (see Figure 1), which itself went through several iterations during that timeframe. Special attention was paid to ensure that the data collected remained the same, so that the dataset remained internally consistent. The dataset was created to study the speech and cognitive markers of individuals with varying levels of cognitive impairment, including dementia. Data were collected during clinical assessments at the Memory Clinic at the Department of Neurology at the Faculty Hospital Královské Vinohrady. Data were collected by a trained psychologist and a speech therapist who conducted all the examinations, from inputting demographic information from patients to guiding the patient through the entire session. The dataset comprises patients with mild cognitive impairments, mild dementia, and a cognitively normal population as a control group. The dataset is published in LINDAT/CLARIN repository [22] and is intended for non-commercial research in clinical neuropsychology, computational linguistics, and machine learning, with the goal of supporting automated detection of cognitive impairment.

FIGURE 4. KOBAR 2 form preview.

Each individual was assessed during a single clinical session. In all cases, only one session per individual is included in the dataset. This approach reduces inter-session variability and reflects the standard memory clinic workflow.

Data for all individuals were checked to ensure that there are no empty or invalid values, and all speech recordings were recognized by an automatic speech recognizer. The fine-tuning was done using in-domain data consisting of, i.e. the nonsense words, the picture names from the sequential picture naming task, and manually transcribed utterances from the complex scene description task.

A. DATA COLLECTION PROCESS

Although the DigiDiaDem application is conceptually designed for future smartphone usage, the current dataset was collected using both laptop and tablet-based devices. Most recordings were obtained in clinical settings using standard internal or external microphones under quiet, controlled conditions. A minority of sessions utilized tablets, which were also supervised in acoustically appropriate clinical rooms. All sessions were conducted by trained professionals (psychologists or speech therapists), ensuring consistency in administration and audio quality.

The examination procedure began with completing the form referred to as KOBAR 2 (see Figure 4). A trained psychologist or a speech therapist entered the individual's sociodemographic information (age, education, sex). Auditory data were then collected using the DigiDiaDem spoken dialog system. Finally, the results of the following cognitive tests and questionnaires were entered:

- The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [38]. RBANS is a standardized neuropsychological battery designed to assess multiple cognitive domains, including immediate and delayed memory, visuospatial/constructional abilities, language, and attention. It consists of twelve subtests that can be administered in about 30 minutes. RBANS assesses five cognitive domains: (1) immediate memory, (2) visuospatial/constructional abilities, (3) language, (4) attention, and (5) delayed memory. The test is widely used to detect cognitive decline and monitor progression in conditions such as dementia and mild cognitive impairment.
- The Mississippi Aphasia Screening Test – Czech version (MASTcz) [39]. The MASTcz is a brief screening tool used to assess language impairment in individuals with aphasia or other neurological disorders. It evaluates expressive and receptive language functions through tasks such as naming, repetition, automatic speech, and comprehension. The Czech adaptation maintains the reliability and validity of the original version, while being culturally and linguistically appropriate for Czech speakers.
- The Amnesia Light and Brief Assessment (ALBA) [28]. The ALBA is a cognitive screening tool focused primarily on the detection of amnestic syndromes and early stages of dementia. It consists of memory-based and orientation tasks that can be completed in a few minutes. It consists of four tasks involving speech (expressive - sentence repetition, and sensory - understanding instructions to demonstrate six gestures) and memory (sentence encoding and recall, and gesture demonstration and recall). It can be completed in two or three minutes without special special aids (e.g., pen or paper) and its brevity and simplicity make it suitable for clinical and research settings requiring cognitive evaluation.
- The Picture Naming and Immediate Recall (PICNIR) [31]. The PICNIR assesses language and episodic memory by asking participants to name pictured objects and recall them. It measures both naming ability and short-term recall, which are often affected in neurodegenerative conditions. The Czech version, abbreviated as POBAV, is adapted to local language and cultural norms, ensuring linguistic validity.
- The Beck Anxiety Inventory (BAI) [40]. The BAI is a self-report questionnaire designed to measure the severity of anxiety symptoms in adults. It includes 21 items describing common physiological and cognitive symptoms of anxiety, rated on a four-point scale. The tool helps distinguish anxiety from depression and is widely used in both clinical and research contexts.
- The Geriatric Depression Scale (GDS) [41]. A screening instrument was developed to assess depressive symptoms in older adults. It consists of simple yes/no questions focused on mood, activity level, and outlook on life, minimizing the influence of physical symptoms. The short form is quick to administer and is particularly suitable for people with mild cognitive impairment or dementia.
- The Functional Activities Questionnaire (FAQ) [42]. The FAQ evaluates a person's ability to perform complex daily activities, such as managing finances, preparing meals, or managing medications. It provides a sensitive indicator of functional decline associated with cognitive impairment. Higher scores indicate greater dependence and are often used to differentiate between normal aging, mild cognitive impairment, and dementia.

When the data were entered, the participants proceeded through the DigiDiaDem application under the supervision of the trained psychologist and a speech therapist. The professionals ensured that the examination went smoothly, but were instructed not to intervene unless necessary. For example, if a participant did not understand the clear task instructions, this was also valuable data for us and could indicate cognitive impairment. Participants were guided through the test battery by TTS instructions that clearly stated what they should do in each task. After the test battery was completed, the professionals could return to the KOBAR 2 form to enter any missing information or to add a note about the participant's behavior or condition.

The entire examination of one patient (or a normal control individual) took approximately 90 minutes, including initial instruction, explanation of individual tests, and their evaluation. The initial introduction and completion of the sociodemographic information took approximately 10 minutes. The DigiDiadem session lasted approximately 10-15 minutes, RBANS took about 30 minutes, ALBA 3 minutes, PICNIR 5 minutes, BAI 2 minutes, FAQ 3 minutes, and GDS took approximately 3 minutes. All participants signed an informed consent. The DigiDiaDem study was approved by the Ethics Committees of the Faculty Hospital Kralovske Vinohrady in Prague, Czech Republic (EK-VP/51/0/2023).

B. DATA CATEGORIZATION

A total number of 371 individuals were assigned into four clinically defined groups:

- **CN** - Community individuals with normal cognition (213 individuals): common people, all of whose RBANS results were in the norm according to international guidelines [43].
- **CMCI** - Community individuals with mild cognitive impairment (70 individuals): people with either total RBANS score or at least one of its subscores lower than the cut-offs [43].
- **MCI** - Patients with mild cognitive impairment (MCI) (46 individuals): with symptoms and objective cognitive impairment based on the judgement of a cognitive neurologist with long-term experience. Most of these patients had been followed at the Memory Clinic after detailed diagnostic work-up, including clinical and neuroimaging evaluation. All the

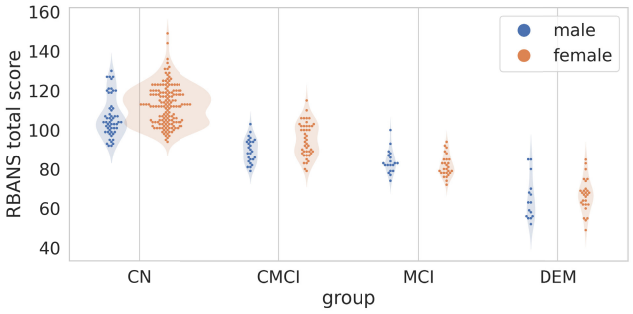


FIGURE 5. Distribution of RBANS total scores across diagnostic groups.

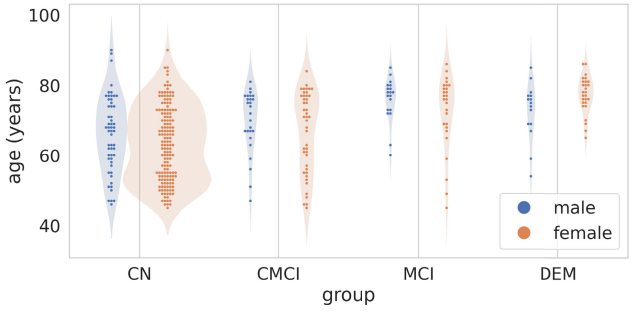


FIGURE 6. Distribution of participants' age across diagnostic groups.

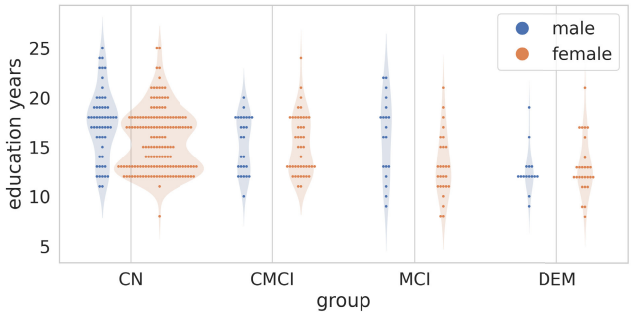


FIGURE 7. Distribution of participants' years of education across diagnostic groups.

cognitive tests and questionnaires in the current study were not used for group categorization since they were collected after an invitation and an agreement to participate in our study by the cognitive neurologist. The MCI patients met criteria of objective cognitive deficit and functional independence required in DSM-5 [43].

- **DEM** - Patients with mild dementia (42 individuals): patients are dependent on caregiver help. The same principle of examination sequence and categorization as for MCI patients applies to these patients with mild dementia.

The clinical labels were established independently by neurologists using standardized instruments, including RBANS and clinical interviews, without reference to DigiDiaDem task scores. To avoid any overlap, DigiDiaDem tasks were administered as a separate module and were not used in the diagnostic decision process. While we did not perform a formal correlation analysis between DigiDiaDem task performance and clinical scores, preliminary observations suggest a partial alignment in cognitively demanding tasks (e.g., verbal fluency), but also highlight an additional signal in speech-derived features not captured by RBANS.

Figures 5, 6, and 7 illustrate the distribution of key demographic and cognitive variables across the four diagnostic groups (CN, CMCI, MCI, and DEM). Overall, the distributions exhibit expected patterns consistent with the clinical progression of cognitive impairment.

In particular, Figure 5 shows the distribution of total RBANS scores, where lower scores correspond to greater levels of impairment. The separation between groups follows the expected gradient, with DEM participants scoring the lowest and CN participants achieving the highest values.

Figure 6 presents the distribution of participant ages. While the groups overlap considerably, there is a mild tendency for older individuals to appear more frequently in the more impaired categories, which is in line with general expectations.

Finally, Figure 7 shows the distribution of years of education. Although the differences between groups are subtle, participants with higher impairment appear to have slightly fewer years of education on average.

Across all figures, there is a clear predominance of female participants. However, this imbalance is not expected to have a meaningful effect on the analysis, as sex is not the primary factor in the examined cognitive outcomes.

C. STRATIFIED TRAIN-TEST PARTITIONING

To enable reproducible model development and unbiased performance evaluation, the dataset was divided into training and testing subsets using a stratified sampling strategy that maintained the proportional representation of key clinical and demographic variables. The stratification key was defined as a composite of three categorical attributes: the final clinical diagnosis, the participant's sex, and the education category. This approach ensured that both subsets preserved the distribution of cognitive conditions, as well as the major socio-demographic characteristics that influence cognitive performance.

Since some combinations of these variables occurred only rarely, a fully stratified split was not always feasible. To address this limitation, an adaptive stratification procedure was implemented. The algorithm first computed the frequency of each unique combination of diagnosis, sex, and education category. Combinations with very few samples were progressively merged into broader categories to ensure that every stratum contained a sufficient number of observations for both the training and testing subsets. The merging process was guided by an automatically determined frequency threshold: the lowest value for which the stratified split could be successfully executed. In the present dataset, this threshold corresponded to four occurrences per unique combination. Below this threshold, rare combinations were replaced by a coarser grouping based solely on participants'

TABLE 1. Summary statistics and proportional distributions for the full dataset, training, and testing subsets after stratified 4:1 split.

		Full dataset	Train set	Test set
Average age	[years]	67.3	67.5	66.5
Partition size		371	294	77
Target class	CN	57.4 %	58.2 %	54.5 %
	CMCI	18.9 %	18.0 %	22.1 %
	MCI	12.4 %	12.6 %	11.7 %
	DEM	11.3 %	11.2 %	11.7 %
Sex	female	69.0 %	69.0 %	68.8 %
	male	31.0 %	31.0 %	31.2 %
Education UNESCO ISCED	5-8	47.7 %	47.6 %	48.1 %
	3A	37.7 %	38.4 %	35.1 %
	3C	12.4 %	11.9 %	14.3 %
	1	2.2 %	2.0 %	2.6 %

sex, thereby preserving sex balance even in underrepresented diagnostic or educational subgroups.

To ensure representativeness of the test subset, the random seed used for the data split was not chosen arbitrarily. Several candidate seeds were tested, and the one yielding the most balanced distribution across diagnostic, demographic, and educational categories was selected (Table 1). Specifically, the chosen seed guaranteed that every target, sex, and education category was represented in the test subset and that the relative proportions of these categories closely matched those observed in the full dataset. In addition to categorical balance, attention was paid to preserving the overall age distribution, ensuring that the mean age and quartiles remained comparable across the training and testing sets. This approach minimizes the risk of systematic bias due to underrepresented strata or age-related effects in the testing data and improves the generalizability of subsequent model evaluation.

The final split allocated approximately 80% of the data to the training set (294 individuals) and 20% to the testing set (77 individuals), corresponding to the intended 4:1 ratio. The stratification procedure achieved near-identical distributions of diagnosis, sex, and education category across the two subsets, with differences remaining within one percentage point for all variables. In addition, the age distribution in those subsets is presented in Table 2. This method ensures that the training and testing sets are both demographically and clinically balanced, thereby providing a robust foundation for the development and evaluation of predictive models of cognitive impairment.

D. AUTOMATIC TRANSCRIPTION OF RECORDINGS

Since our goal is to develop an automated method for early detection of cognitive disorders, we need to employ automatic speech recognition to transcribe the recorded speech signal into a corresponding text representation. Such transcription is challenging due to multiple factors — technically induced distortions (e.g., background noise, far-field

TABLE 2. Mean age [years] in different data partitions vs. assigned target class.

		Full dataset	Train set	Test set
Target class	CN	64	64	62
	CMCI	68	68	68
	MCI	74	74	74
	DEM	76	75	78

recording conditions, or voice-over-IP transmission), natural speech phenomena (such as hesitations and repetitions), and task-related characteristics of the participants, who are often elderly and may already exhibit early signs of dementia.

In our prior work [19], we evaluated the performance of a zero-shot Wav2Vec 2.0 speech recognition model [44]. Patients and control individuals were asked to describe a black-and-white vivid shore scene, which served as the first version of the stimulus shown in Figure 2. To establish a reliable ground truth, all recorded utterances were manually transcribed. Subsequently, the recognized utterances generated by the ASR system were evaluated using word- and character-level accuracy metrics, computed with the Levenshtein distance. The results demonstrated solid recognition accuracy for both the patients and the control individuals. The ASR reached 90% character-level accuracy in both groups, and word-level accuracies of 67.4% for the patient group and 74.1% for the cognitively normal control group. These encouraging findings provided the confidence to proceed with the project.

The next step was to improve the ASR accuracy. For this dataset, we transcribed every complex scene description recording and used the transcriptions to fill the language model of the ASR. For the transcription of speech recordings, we used the software *Transcriber* [45]. It allows precise alignment of spoken content with time-coded markers. In addition to transcribing the speech itself, we also annotated a range of acoustic and contextual events:

- background noise (e.g., paper rustling, environmental sounds),
- TTS prompts (segments spoken by the text-to-speech system),
- pauses and silences,
- overlapping speech,
- unfinished words (denoted with "+" - unfinished word "*animal*" would be "*anima+*" or "*+nimal*")
- non-verbal sounds (e.g., laughter, sighs, coughing),
- emphasis or marked intonation.

The result was a text representation of the recording with timestamps, enabling accurate navigation within the original audio.

We also took a list of 1315 of the most common animal names that were then used to fill the language model.

We considered two different neural architectures (Wav2Vec 2.0 and Zipformer) in combination with different language

TABLE 3. Comparison of ASR models (character and word-level accuracies).

ASR model	char Acc	word Acc
wav2vec_no_lm	86.1	68.3
wav2vec_generic	86.1	75.4
wav2vec_lm-extra06	88.9	84.6
zipformer	89.1	83.7
zipformer_lm-extra06	94.4	89.8

models. In total, five ASR models were evaluated (see Tab. 3) on test data:

- wav2vec_nolm: A baseline recognizer built on the Wav2Vec 2.0 base encoder, which was used in the preliminary experiments. Used as a reference.
- wav2vec_generic: Uses the same Wav2Vec 2.0 acoustic network Transformer as wav2vec_no_lm, but its hypotheses are rescored with a large back-off 5-gram LM trained on billions of Czech web text words. The language model improves recognition performance while leaving acoustic parameters unchanged.
- wav2vec_lm-extra06: Shares the same acoustic front-end as the previous models but employs a smaller 5-gram language model fine-tuned on transcripts from the in-domain dataset. This tight domain adaptation allows the recognizer to favour in-vocabulary words and phrases.
- zipformer: Trained from scratch on the same Czech speech corpora using the Zipformer encoder [46]. It employs U-Net-style subsampling and BiasNorm layers, resulting in faster inference and better handling of disfluent or phonetically irregular speech. It produces SentencePiece tokens via greedy CTC decoding, functioning as a language-model-free alternative to wav2vec_no_lm.
- zipformer_lm-extra06: Same architecture as zipformer, only enhanced with a LM with in-domain data.

There is a significant 16.3% improvement in word-level accuracy when using the Wav2Vec 2.0 model with an in-domain language model (wav2vec_lm-extra06) compared to without any language model (wav2vec_no_lm).

Interestingly, the Zipformer (zipformer) encoder, despite not using a language model, attains nearly the same word-level accuracy (83.7% vs 84.6%) and even higher character-level accuracy (89.1% vs 88.9%) than Wav2Vec 2.0 with an in-domain language model. This can be attributed to Zipformer's architecture, which efficiently captures long-range dependencies and context directly from the acoustic signal. In contrast, Wav2Vec 2.0, while highly effective, relies more on a language model to incorporate domain-specific lexical and syntactic knowledge, highlighting the trade-off between acoustic modelling strength and language model dependence.

TABLE 4. Number of features for each task.

Task	01	02	03	04m	04s	05	06	07	08	09	10
# features	1	1	2	6	9	3	2	1	3	2	1

There is also an improvement when using the Zipformer model with a language model (zipformer_lm-extra06) compared to the Zipformer encoder without one (zipformer). The improvement is 5.3% for character-level accuracy and 6.1% for word-level accuracy, confirming the usefulness of an in-domain language model. zipformer_lm-extra06 was subsequently used for the automatic speech recognition of all recordings.

E. MANUALLY ENGINEERED FEATURES

A set of baseline, manually engineered, linguistic and acoustic features was designed for each task, expanding on the cognitive goals presented in the previous section.

It is worth noting that not all recordings are the same length for each task. That is because we apply a dynamic timeout extension for some of the tasks (namely, tasks 01, 02, 03, 06, 09, 10). The timeouts for automatic speech recognition are set to either 5 or 10 seconds, and we did not want to lose any speech data if the participants started speaking after a delay (cognitively impaired participants might take longer to understand the prompt of the task), so we extended the timeout in that particular instance. A possible counterargument to the dynamic timeout extension is that the timeout could simply be set to a longer duration for all participants. However, we chose not to do this, as we wanted to avoid having participants who respond quickly spend more time on the examination than necessary.

From the recognized audio transcriptions, a set of manually engineered features was extracted for each task. The primary goal was to design a collection of simple, interpretable features that could be computed directly from the participants' responses, that is, from their automatically recognized speech. These features were intended to capture relevant linguistic and acoustic patterns while remaining computationally efficient and easily applicable in practical screening or assessment scenarios.

A total of 31 manually engineered features were used for classification, with each task represented by a different number of features as detailed in Tab. 4. The extracted features include a mix of speech-derived metrics (e.g., acoustic duration, word entropy), task-specific behavioral scores (e.g., number of correct items, semantic loss), and interaction-based features (e.g., latency, repetitions).

1) TASK 01: HEARING EXAMINATION

- *Correctly repeated numbers* quantifies the participant's ability to correctly repeat a sequence of numbers as presented during the hearing test. It reflects hearing, short-term auditory memory and auditory speech perception.

2) TASK 02: VISION EXAMINATION

• *Correctly repeated characters* measures the participant's capacity to visually recognize and accurately repeat displayed characters (letters or digits). It indexes eyesight, short-term visual memory and attention.

3) TASK 03: SENTENCE REPETITION

• *Character match ratio* evaluates the proportion of characters that match between the model sentence and the participant's repetition, excluding spaces. It expresses the literal similarity of the reproduced sentence to the target.

• *Correctly repeated words* represents the number of words from the model sentence that were repeated correctly in the participant's response. It captures lexical recall and articulation accuracy.

4) TASK 04: COMPLEX SCENE DESCRIPTION

For this task, the features were generated using two different approaches: (1) by morphological analysis using MorphoDiTa [47], and (2) by matching machine readable reference with a subject's complex scene description [20]. Since the approaches are fundamentally different, we treat them as different feature sets:

a: TASK 04m: MORPHODITA BASED FEATURES

Morphological analysis was used to extract detailed linguistic characteristics of the scene description. By applying part-of-speech tagging and lemmatization, we calculated the counts and relative proportions of the major word classes. Additional text-level features were derived, such as the number of sentences and ratios between selected word categories.

• *Sentence count* measures the total number of sentences produced during the one-minute spontaneous description of the complex picture. It indexes verbal productivity and narrative fluency.

• *First person verb proportion* expresses the proportion of verbs produced in the first person singular form (e.g., "*I see*," "*I know*") relative to all verbs in the description. It captures self-referential language use.

• *Meaningful words ratio* quantifies the share of content words (e.g., nouns and main verbs) among all produced tokens. It reflects lexical informativeness.

• *Pronoun to noun ratio* measures the relative frequency of pronouns (e.g., "*he*," "*it*," "*I*") compared with nouns. Higher values indicate pronoun-heavy, more anaphoric discourse; lower values indicate noun-rich, more referential discourse.

• *Count of repeated meaningful words* counts how many times content lemmas (e.g., core noun/verb lexemes) are repeated across the description. It indexes redundancy and potential perseveration.

• *Unique words to total words* captures lexical diversity as the ratio of distinct word types to total tokens in the description. Higher values indicate more varied vocabulary.

b: TASK 04s: COMPLEX SCENE DESCRIPTION FEATURES

Using custom semantic-parsing pipeline and leveraging a richly annotated reference description, our system extracts interpretable semantic loss vectors that quantify deviations from an expert-defined ground truth.

• *Named object count* is the count of distinct visual objects in the scene that are explicitly named in the narrative. It reflects referential coverage of the scene.

• *Described object relation count* counts the number of explicit relations stated between objects (e.g., spatial, functional, or action relations). It indexes relational and syntactic expressiveness.

• *Distinct topic count* measures how many different topical foci are introduced in the description (e.g., subevents or thematic clusters). It reflects discourse organization.

• *Description trajectory length* quantifies the cumulative spatial "path length" implied by the order in which objects are mentioned, computed as the total distance between consecutively referenced objects in the picture. It reflects the scanning strategy and narrative traversal of the scene.

• *Objects in water count* counts the objects explicitly mentioned as being in the water region of the scene. It indicates coverage of that visual sub-area.

• *Objects in sky count* counts the objects explicitly mentioned as being in the sky region of the scene. It indicates coverage of that visual sub-area.

• *Objects on land count* is the number of objects explicitly mentioned as being on land in the scene. It indicates coverage of that visual sub-area.

• *Explicit child danger mentioned* is a binary indicator of whether the narrative contains an explicit statement of danger involving the child in the scene.

• *Explicit animal danger mentioned* is a binary indicator of whether the narrative contains an explicit statement of danger involving the dog or cat in the scene.

5) TASK 05: COMPLEX SCENE OBJECT RECALL

• *Total recalled words count* reflects the total number of words produced when recalling objects from the previously described scene within the time limit. It indexes recall performance.

• *Distinct objects recalled count* evaluates the distinct scene objects correctly recalled without duplication. It reflects recall accuracy and breadth.

• *Repeated recalled words count* measures the count of words that were uttered multiple times during the recall task, indexing repetition and monitoring.

6) TASK 06: SEQUENTIAL PICTURE NAMING

• *Correctly named pictures count* is the number of pictures (out of 20) that were correctly named during sequential presentation. It reflects object naming ability.

• *Total naming reaction time* sums the latencies from picture onset to speech onset across all trials, capturing global naming speed. If a participant fails to name a picture correctly,

a penalty of 10 seconds is assigned as the reaction time for that trial.

7) TASK 07: PICTURE RECALL

• *Correctly recalled pictures count* is the number of the black-and-white pictures correctly recalled after the naming phase. It indexes short-term visual memory for pictured objects.

8) TASK 08: ANIMAL VERBAL FLUENCY

• *Total word count* represents the total number of words produced while naming animals within the time constraint. It reflects psychomotor speed, semantic fluency and retrieval speed.

• *Animal word count* is the number of unique animal names produced without repetition, indexing semantic access and lexical diversity.

• *Repeated animals count* evaluates the animal names that were repeated within the task, reflecting inhibitory control and self-monitoring.

9) TASK 09: SENTENCE RECALL

• *Percentage of repeated-recalled sentence characters* evaluates the proportion of matching characters between the recalled sentence and the target sentence. It reflects fine-grained recall fidelity.

• *Correct recalled sentence words count* measures how many words from the target sentence were recalled correctly. It captures lexical recall quality.

10) TASK 10: NONSENSE WORDS REPETITION

• *Word similarity score* reports the mean F1 score based on Levenshtein alignment between the produced nonsense words and their targets, capturing phonological reproduction accuracy for pseudowords. Since data collection for this task started later and the set of target pseudowords changed over time, the results were computed only for the intersection of six words common to both phases.

F. SOCIODEMOGRAPHIC FEATURES

Basic sociodemographic features were also incorporated into the dataset to control for their potential impact on cognitive performance.

• *Age* represents the participant's chronological age in years. It serves as a key covariate due to its strong association with cognitive performance and age-related decline.

• *Sex* denotes the participant's sex (coded as male or female). It is included to control for potential sex-related differences in cognitive and linguistic performance.

• *Years of education* indicates the total number of years of formal education completed by the participant. It captures educational attainment, a factor known to influence cognitive reserve and task performance.

IV. EXPERIMENTS

A large number of experiments can be performed over a set of four target classes. To define the baseline experiments, we present only a binary classification task separating community individuals from patients with mild cognitive impairment or with mild dementia:

TABLE 5. Classification results (LR) for two-class (CN vs. MCI+DEM) classifier trained on train and evaluated on test partition respectively.

task	targets	precision	recall	F1-score	accuracy
01	CN	0.700	1.000	0.824	0.700
	MCI+DEM	0.000	0.000	0.000	
02	CN	0.700	1.000	0.824	0.700
	MCI+DEM	0.000	0.000	0.000	
03	CN	0.787	0.881	0.831	0.750
	MCI+DEM	0.615	0.444	0.516	
04m	CN	0.783	0.857	0.818	0.733
	MCI+DEM	0.571	0.444	0.500	
04s	CN	0.814	0.833	0.824	0.750
	MCI+DEM	0.588	0.556	0.571	
05	CN	0.851	0.952	0.899	0.850
	MCI+DEM	0.846	0.611	0.710	
06	CN	0.792	0.905	0.844	0.767
	MCI+DEM	0.667	0.444	0.533	
07	CN	0.884	0.905	0.894	0.850
	MCI+DEM	0.765	0.722	0.743	
08	CN	0.927	0.905	0.916	0.883
	MCI+DEM	0.789	0.833	0.811	
09	CN	0.884	0.905	0.894	0.850
	MCI+DEM	0.765	0.722	0.743	
10	CN	0.780	0.929	0.848	0.767
	MCI+DEM	0.700	0.389	0.500	

TABLE 6. Classification results (LR) for two-class (CN vs. MCI+DEM) classifier evaluated using 5-fold cross-validation over the whole dataset.

task	targets	precision	recall	F1-score	accuracy
01	CN	0.708	1.000	0.829	0.708
	MCI+DEM	0.000	0.000	0.000	
02	CN	0.708	1.000	0.829	0.708
	MCI+DEM	0.000	0.000	0.000	
03	CN	0.780	0.930	0.848	0.764
	MCI+DEM	0.681	0.364	0.474	
04m	CN	0.794	0.925	0.855	0.777
	MCI+DEM	0.698	0.420	0.525	
04s	CN	0.812	0.892	0.850	0.777
	MCI+DEM	0.652	0.494	0.562	
05	CN	0.879	0.925	0.902	0.857
	MCI+DEM	0.792	0.693	0.739	
06	CN	0.788	0.925	0.851	0.771
	MCI+DEM	0.686	0.398	0.504	
07	CN	0.878	0.911	0.894	0.847
	MCI+DEM	0.762	0.693	0.726	
08	CN	0.900	0.925	0.912	0.874
	MCI+DEM	0.805	0.750	0.776	
09	CN	0.903	0.915	0.909	0.870
	MCI+DEM	0.788	0.761	0.775	
10	CN	0.758	0.933	0.836	0.739
	MCI+DEM	0.609	0.259	0.364	

- **CN**: Community individuals with normal cognition.
- **MCI + DEM**: Patients with mild cognitive impairment or diagnosed mild dementia.

Alternative binary groupings of the four classes, together with the corresponding multi-class formulation, are provided in a companion script available in the Github repository `github.com/SALT-UWB/diadem`. For the four-class classification setup (CN vs sMCI vs mMCI vs DEM), the Histogram Gradient Boosting Trees (HGBT) classifier was used in its native multiclass mode without class weighting. HGBT was utilized as a high-performance non-linear baseline to complement traditional Logistic Regression. No oversampling or class-balancing techniques were applied, as our goal was to establish baseline performance under natural clinical data distributions. The script includes a dataset preview, illustrates the preprocessing workflow, and reproduces all experiments reported in this paper.

All classification experiments (e.g., logistic regression, HGBT) and speech recognition were run on a CPU without GPU acceleration. ASR-related model fine-tuning (e.g., Wav2Vec 2.0, Zipformer) was performed using GPU nodes within the Czech national computing infrastructure, MetaCentrum.

A. BASELINE EXPERIMENTS

The baseline experiments conducted in this section provide a generic, lower bound on classification performance over different tasks and corresponding feature sets presented in the previous sections. The classifiers were not optimized for maximum performance; instead, we employed a straightforward pipeline consisting of basic preprocessing (mean imputation followed by standardization) and a logistic regression classifier with default hyperparameter settings, as available in scikit-learn's `LogisticRegression` implementation. For the baseline experiments presented in this paper, each cognitive task was modeled independently. Features were extracted and classified on a per-task basis to evaluate the diagnostic potential of individual sub-tests without considering cross-task interactions. For each task and binary classification problem (classes CN vs. MCI+DEM), we always reported precision, recall, F1 score and overall classification accuracy. We trained the models on the train partition and evaluated the metrics on the test partition of the DigiDiadem dataset (see Tab. 5). To validate the correctness of the stratification, we also evaluated classification performance using 5-fold cross-validation on the entire dataset (Tab. 6).

Tasks 01 (Hearing examination) and 02 (Vision examination) showed zero precision and recall for the MCI+DEM class. This outcome was expected, as these tasks were not designed to discriminate between cognitively normal participants and those with cognitive impairment. Their simplicity is intentional: they serve as preliminary screening steps to ensure that participants are able to hear the auditory prompts and visually perceive on-screen stimuli, respectively. In particular, Task 02 involves a brief recognition of a static alphanumeric string and is not intended to provide cognitive challenge or diagnostic value. Consequently, the model correctly identified all CN cases but completely missed MCI+DEM instances in these two tasks, as they lack discriminative features by design.

Starting from Task 03 (Sentence repetition), the model began to detect MCI+DEM participants with moderate performance (precision 0.615, recall 0.444), while maintaining high accuracy for CN (F1-score 0.831). Tasks 04m and 04s (Complex scene description) showed similar or slightly improved results for MCI+DEM, indicating a gradual increase in the model's ability to recognize cognitive impairment. Task 05 (Complex scene object recall) demonstrated strong results for both groups, achieving high F1-scores for CN (0.899) and MCI+DEM (0.710). Task 06 (Sequential picture naming) maintained comparable performance, though the recall for MCI+DEM is slightly decreased. Tasks 07 (Picture recall), 08 (Animal verbal fluency), and 09 (Sentence recall) represented the best-performing group of tasks, showing a balanced trade-off between both classes. In particular, Task 08 stood out as the overall best, with high F1-scores for both CN (0.916) and MCI+DEM (0.811), suggesting effective discrimination between cognitively normal and impaired participants. Finally, Task 10 (Nonsense words repetition) retained strong performance for CN (F1-score 0.848) but showed a moderate drop in recall for MCI+DEM (0.389), which slightly reduced its F1-score (0.500).

Overall, these results suggest that while early sensory tasks (01–02) did not contribute to cognitive status discrimination, as expected, later and more cognitively demanding tasks (05–09) provided the most informative features for distinguishing between CN and MCI+DEM participants.

In addition to logistic regression, we also evaluated a tree-based HGBT classifier provided in the GitHub companion script. Its performance was slightly higher, but the overall trends remained consistent across models.

B. ADDITION OF SOCIODEMOGRAPHIC FEATURES

Given that the risk of cognitive impairment increases with age and lower education, sociodemographic features - age, sex, and years of education - were added among the classification features for each individual task. The results are summarized in Table 7 and Table 8. Extending the classification features by sociodemographic information led to a consistent improvement. This aligns with expectations, as these variables are known to correlate with cognitive decline.

While age and education are established risk factors, sex is not considered a primary predictor of cognitive impairment in our model. Nevertheless, sex is an important feature in voice-based classification because it systematically shapes the speech signal and the produced content. Beyond differences in voice pitch and vocal-tract acoustics, sex influences phonetic patterns and – crucially in Czech –

TABLE 7. Classification results (LR) for two-class (CN vs. MCI+DEM) classifier trained on train and evaluated on test partition respectively INCLUDING sociodemographic features.

task	targets	precision	recall	F1-score	accuracy
01	CN	0.851	0.952	0.899	0.850
	MCI+DEM	0.846	0.611	0.710	
02	CN	0.816	0.952	0.879	0.817
	MCI+DEM	0.818	0.500	0.621	
03	CN	0.822	0.881	0.851	0.783
	MCI+DEM	0.667	0.556	0.606	
04m	CN	0.872	0.976	0.921	0.883
	MCI+DEM	0.923	0.667	0.774	
04s	CN	0.867	0.929	0.897	0.850
	MCI+DEM	0.800	0.667	0.727	
05	CN	0.933	1.000	0.966	0.950
	MCI+DEM	1.000	0.833	0.909	
06	CN	0.841	0.881	0.860	0.800
	MCI+DEM	0.688	0.611	0.647	
07	CN	0.886	0.929	0.907	0.867
	MCI+DEM	0.812	0.722	0.765	
08	CN	0.927	0.905	0.916	0.883
	MCI+DEM	0.789	0.833	0.811	
09	CN	0.889	0.952	0.920	0.883
	MCI+DEM	0.867	0.722	0.788	
10	CN	0.870	0.952	0.909	0.867
	MCI+DEM	0.857	0.667	0.750	

TABLE 8. Classification results (LR) for two-class (CN vs. MCI+DEM) classifier evaluated using 5-fold cross-validation over the whole dataset INCLUDING sociodemographic features.

task	targets	precision	recall	F1-score	accuracy
01	CN	0.820	0.878	0.848	0.777
	MCI+DEM	0.644	0.534	0.584	
02	CN	0.823	0.873	0.847	0.777
	MCI+DEM	0.640	0.545	0.589	
03	CN	0.822	0.869	0.845	0.774
	MCI+DEM	0.632	0.545	0.585	
04m	CN	0.837	0.915	0.874	0.814
	MCI+DEM	0.735	0.568	0.641	
04s	CN	0.834	0.897	0.864	0.801
	MCI+DEM	0.694	0.568	0.625	
05	CN	0.905	0.944	0.924	0.890
	MCI+DEM	0.848	0.761	0.802	
06	CN	0.850	0.906	0.877	0.821
	MCI+DEM	0.730	0.614	0.667	
07	CN	0.887	0.920	0.903	0.860
	MCI+DEM	0.787	0.716	0.750	
08	CN	0.895	0.925	0.910	0.870
	MCI+DEM	0.802	0.739	0.769	
09	CN	0.913	0.934	0.923	0.890
	MCI+DEM	0.831	0.784	0.807	
10	CN	0.815	0.869	0.841	0.767
	MCI+DEM	0.622	0.523	0.568	

morphological and semantic properties of utterances due to pervasive grammatical-gender marking in a highly inflectional language. Including sex therefore helps the model account for predictable linguistic variation and improves the robustness of feature extraction.

Although our dataset contains a higher proportion of female participants (consistent with the demographic profile of the memory clinic population), sex is included to control for systematic sex-dependent variation in speech and language, and to support reliable comparisons across speakers. This supports the interpretation of sex not as a disease-related factor, but as a necessary adjustment variable when working with voice-based models.

In particular, the inclusion of sociodemographic features substantially increased classification performance in early tasks, which previously lacked discriminatory power.

More complex tasks, such as 04m and 04s, also benefited considerably. This enhancement indicates that the combined effect of task-specific performance and demographic context provides a more comprehensive representation of participants' cognitive state. Later tasks (07–10) already performed well in the baseline model, but modest further improvements were still observed.

Overall, the integration of sociodemographic variables enhanced performance in previously weaker tasks and slightly improved the results across the full set, suggesting that these features provide complementary information that reinforces the model's generalizability.

V. DISCUSSION

Our findings indicate that digital cognitive testing combined with automated speech analysis can effectively detect cognitive impairment. Even with straightforward models, we achieved high accuracy in distinguishing cognitively impaired participants from community individuals with normal cognition.

Moreover, incorporating sociodemographic factors (age, education, etc.) further improved the detection accuracy (up to 95% in our experiments), indicating that personal characteristics provide complementary information for assessing dementia risk. These results are in line with recent studies that leverage speech-based and telemedical assessments for early dementia detection [11], [14] and they support the notion that well-designed remote cognitive tests can complement traditional in-person assessments by enabling preliminary, at-home screening and identifying individuals who may benefit from further neurological evaluation.

Despite these promising results, several limitations and open questions remain. First, our dataset and experiments are specific to Czech-speaking participants, so it is uncertain how well the approach generalizes to other languages or clinical settings. This language/cultural specificity may limit the immediate applicability of our model internationally, an issue that future studies must address by testing and adapting the approach for diverse populations. Second, in this initial work we focused on simple, interpretable classifiers (logistic regression and HGBT) as baselines. It is possible

that more complex non-linear models (e.g., neural networks) could yield higher accuracy. We chose not to include such models here due to space and scope, but exploring them is an important direction for further research, especially to see if they provide significant gains over linear models and how to balance those gains with interpretability. Third, we have not yet performed a detailed analysis of feature importance or ablation to determine which specific tasks or speech-derived features contribute most to the classification outcomes. As a result, while we observed that certain tasks (e.g., Task 08 Animal verbal fluency, Task 05 Complex scene object recall) were particularly informative, we cannot conclusively rank the predictors of cognitive impairment. Future work will include systematic ablation experiments to evaluate the contribution of each task individually, as well as comparative models trained on only speech features, only task performance scores, and combined inputs. This will help isolate the relative impact of each input modality and improve interpretability. Understanding the relative importance of each feature is crucial for clinical interpretability, and we acknowledge this as a priority for future work.

Building on the above limitations, we identify several open research questions and future directions for this line of research:

Generalizability across languages and settings: How well can digital speech-cognitive screening tools like DigiDiaDem perform in other languages or healthcare contexts? Our current results are based on Czech data; validating and calibrating the approach for different languages and cultural backgrounds is essential. While the dataset is Czech-specific, it reflects the clinical reality of underrepresented linguistic populations. Some speech and task-derived features (e.g., fluency, hesitation patterns) may generalize across languages, but further work is needed to validate cross-lingual transferability. This remains an open research question. Future work should include cross-lingual studies or the creation of similar datasets in other populations to ensure broader applicability of the method.

Feature importance and interpretability: Which specific tasks and features are the most indicative of cognitive impairment? We plan to investigate this by conducting detailed feature importance analyses and ablation studies. For example, task scores could be normalized (e.g., converted to z-scores) to make feature contributions comparable, and methods like coefficient analysis or tree-based feature importance can identify the strongest predictors. Such insights will not only answer scientific questions about dementia markers but also enhance the interpretability of the model's decisions for clinical use.

Advanced modeling vs. simplicity: Another open question is whether more complex machine learning models would substantially improve detection performance. While our baseline logistic regression proved surprisingly effective, exploring non-linear classifiers (such as Random Forests or neural networks) in future experiments could reveal additional performance gains. Any improvement must be weighed against the need for transparency – an important consideration in medical AI. Investigating this trade-off is crucial: how to achieve the best accuracy without sacrificing interpretability remains a key research challenge.

Multimodal contributions: We also need to understand how different input modalities contribute to the overall assessment. Our application combines cognitive task performance data with speech-derived features; however, the individual contribution of each modality (cognitive vs. speech) has yet to be quantified. Future studies should consider separate evaluations (e.g., using only cognitive test scores vs. only speech features) and combined analyses to determine the added value of each modality. Additionally, techniques such as task weighting or attention mechanisms could be explored to automatically learn which tasks are most informative for the final decision.

By addressing these open questions, future research can build upon our initial findings and the open, publicly available dataset to develop a more robust, generalizable, and interpretable system for early dementia detection. In particular, our immediate next steps will focus on feature importance analysis and model enhancements as noted above.

Dementia diagnosis and care inherently require a multidisciplinary approach, involving neurologists, psychologists, speech-language therapists, and other specialists. While DigiDiaDem focuses on the speech and cognitive interface, it is not intended to replace comprehensive clinical evaluation. Rather, it complements existing workflows and enables scalable pre-screening that can support expert teams in triaging, monitoring, or referring patients more effectively.

VI. CONCLUSION

The DigiDiaDem application was developed to collect a comprehensive dataset from individuals with varying levels of cognitive functions. The dataset contains a basic set of core demographic covariates (age, education, sex), results of in-person assessment tests (RBANS, ALBA, PICNIR, MASTcz), and data taken from the DigiDiaDem application focused on memory, naming, fluency, description, and repetition. Participants were assigned into four groups based on clinical assessment: the community individuals with normal cognition, the community individuals with mild cognitive impairment, the patients with mild cognitive impairment, and the patients with mild dementia. All recordings and examinations were conducted as additional assessments to routine clinical practice in patients and in community individuals at the Department of Neurology. Participants were assessed using in-person cognitive tests and questionnaires. Furthermore, they engaged in a dialog with the DigiDiaDem application. The dataset, which contains 371 examinations, was divided into a training and test part using stratification by clinical group, age, sex, and level of education to ensure

an even distribution of these key characteristics in both parts of the data.

Even the simple classifier based on logistic regression achieves a relatively high F1-score in mild cognitive impairment / dementia detection, reaching the best result of 0.811 for Task 08 (Animal verbal fluency) in the traditional train/test machine-learning setup. In this case, the accuracy reaches 88.3% (Tab. 5). The inclusion of sociodemographic features adds important information, and the accuracy improves to 95% (Task 05, Complex scene object recall) with F1-score 0.909 (Tab. 6).

The aim of the described – open and publicly available – dataset is to support the development of cognitive disorders detection methods based on automatic speech analysis and cognitive performance assessment. The dataset is suitable for research in the areas of clinical neuropsychology, computational linguistics, and machine learning. The DigiDiaDem dataset is published together with train/test partitioning, sociodemographic features, manually engineered features for each task and automatic speech transcriptions.

The paper presents the manually engineered features defined for each subtask of the DigiDiaDem examination and subsequently performs a baseline classifier evaluation using logistic regression. Such experiments are valuable for the reproducibility of other experiments conducted using the DigiDiaDem dataset.

Our results demonstrate the feasibility of using well-designed telemedical cognitive tests for classification tasks, potentially substituting certain traditional in-person expert-administered assessments.

REFERENCES

[1] R. Priester, R. L. Kane, and A. M. Totten. *How the Current System Fails People With Chronic Illness*. Accessed: Aug. 20, 2025. [Online]. Available: https://www.soa.org/globalassets/assets/files/resources/essays-monographs/financing-chronic-care/m-hb05-1vi.pdf

[2] A. Kalache and K. Sen. (Aug. 2017). *Ageing and Health*. World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[3] G. Livingston et al., "Dementia prevention, intervention, and care," *Lancet*, vol. 390, no. 10113, pp. 2673–2734, Dec. 2017.

[4] C. M. Callahan, T. Foroud, A. J. Saykin, A. Shekhar, and H. C. Hendrie, "Translational research on aging: Clinical epidemiology as a bridge between the sciences," *Transl. Res.*, vol. 163, no. 5, pp. 439–445, May 2014, doi: 10.1016/j.trsl.2013.09.002.

[5] World Health Organization. (2025). *Dementia*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[6] H. Brodaty, D. Pond, N. M. Kemp, G. Luscombe, L. Harding, K. Berman, and F. A. Huppert, "The GPCOG: A new screening test for dementia designed for general practice," *J. Amer. Geriatrics Soc.*, vol. 50, no. 3, pp. 530–534, Mar. 2002, doi: 10.1046/j.1532-5415.2002.50122.x.

[7] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra, "Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks," *J. Clin. Experim. Neuropsychol.*, vol. 40, no. 9, pp. 917–939, Oct. 2018, doi: 10.1080/13803395.2018.1446513.

[8] H. Goodglass and E. Kaplan, "The cookie theft picture," in *Boston Diagnostic Aphasia Examination*. Boston, MA, USA: Veterans Administration Hospital, 1972.

[9] 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 Lang.*, vol. 53, no. 1, pp. 1–19, Apr. 1996, doi: 10.1006/brln.1996.0033.

[10] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task," *Neurological Sci.*, vol. 26, no. 4, pp. 243–254, Oct. 2005, doi: 10.1007/s10072-005-0467-9.

[11] E. Fristed, C. Skirrow, M. Meszaros, R. Lenain, U. Meepegama, S. Cappa, D. Aarsland, and J. Weston, "A remote speech-based AI system to screen for early Alzheimer's disease via smartphones," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitor.*, vol. 14, no. 1, pp. 1–14, Jan. 2022, doi: 10.1002/dad2.12366.

[12] F. de Arriba-Pérez, S. García-Méndez, F. J. González-Castaño, and E. Costa-Montenegro, "Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with natural language processing capabilities," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 12, pp. 16283–16298, Dec. 2023, doi: 10.1007/s12652-022-03849-2.

[13] I. S. Beschi, K. Helenprabha, S. M. Krishna, A. Deepa, S. Murugan, and S. B. T. Naganathan, "Advancing remote cognitive assessments through telemedicine with cloud computing and AdaBoost classifiers," in *Proc. 11th Int. Conf. Commun. Signal Process. (ICCSP)*, Melmaruvathur, India, Jun. 2025, pp. 458–463, doi: 10.1109/iccsp64183.2025.11089139.

[14] A. König, R. Zeghari, R. Guerchouche, F. Bremond, V. Bultingaire, P. Robert, and N. Linz, "Remote cognitive assessment of older adults in rural areas by telemedicine and automatic speech and video analysis," *Alzheimer's Dementia*, vol. 17, no. S8, pp. 1–10, Dec. 2021, doi: 10.1002/alz.050302.

[15] J. Robin, M. Xu, L. D. Kaufman, and W. Simpson, "Using digital speech assessments to detect early signs of cognitive impairment," *Frontiers Digit. Health*, vol. 3, pp. 1–9, Oct. 2021, doi: 10.3389/fdgth.2021.749758.

[16] M. Nunes, A. S. Teles, D. Farias, C. Diniz, V. H. Bastos, and S. Teixeira, "A telemedicine platform for aphasia: Protocol for a development and usability study," *JMIR Res. Protocols*, vol. 11, no. 11, Nov. 2022, Art. no. e40603, doi: 10.2196/40603.

[17] J. Cohen-Mansfield, "Cohen-mansfield agitation inventory (CMAI) [Database record]," APA PsycTests, Amer. Psychol. Assoc., Washington, DC, USA, Tech. Rep. t28576-000, 1986, doi: 10.1037/t28576-000.

[18] J. Rosen, L. Burgio, M. Kollar, M. Cain, M. Allison, M. Fogleman, M. Michael, and G. S. Zubenko, "The Pittsburgh agitation scale: A user-friendly instrument for rating agitation in dementia patients," *Amer. J. Geriatric Psychiatry*, vol. 2, no. 1, pp. 52–59, Dec. 1994, doi: 10.1097/00019442-199400210-00008.

[19] J. Švec, F. Polák, A. Bartoš, M. Zapletalová, and M. Víta, "Evaluation of Wav2Vec speech recognition for speakers with cognitive disorders," in *Proc. Int. Conf. Text, Speech, Dialogue*, vol. 13502. Cham, Switzerland: Springer, 2022, pp. 501–512, doi: 10.1007/978-3-031-16270-1_41.

[20] T. Lebeda, L. Zajícová, J. Švec, and L. Šmídl, "Automatic cognitive disorder detection through semantic analysis of verbal image descriptions," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, Aug. 2025, pp. 177–189, doi: 10.1007/978-3-032-02551-7_16.

[21] J. Tupý, J. Švec, and L. Šmídl, "Detection of cognitive disorders using ASR-based nonsense words repetition," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, Aug. 2025, pp. 247–258, doi: 10.1007/978-3-032-02548-7_21.

[22] L. Šmídl, M. Kompasová, M. Zapletalová, F. Polák, L. Zajícová, J. Švec, M. Víta, and A. Bartoš. (2025). *DigiDiaDem Speech-Cognitive Dataset (DSCD-CZ-2)*. [Online]. Available: http://hdl.handle.net/11234/1-6043

[23] E. F. Kaplan, H. Goodglass, and S. Weintraub, *The Boston Naming Test*. Philadelphia, PA, USA: Lea & Febige, 1983.

[24] M. Michalovová and A. Bartos, "Brief cognitive tests for clinical practice," *Ceskáa Slovenská Neurologie Neurochirurgie*, vol. 88, no. 4, pp. 210–224, Aug. 2025, doi: 10.48095/cccsnn2025210.

[25] A. Bartoš, "Brief test of verbal memory using the sentence in Alzheimer disease," *Cesk Slov Neurol.*, vol. 80, no. 6, pp. 679–684, Nov. 2017, doi: 10.14735/amcsnn2017679.

[26] C. T. Chun, K. Seward, A. Patterson, A. Melton, and L. Macdonald-Wicks, "Evaluation of available cognitive tools used to measure mild cognitive decline: A scoping review," *Nutrients*, vol. 13, no. 11, p. 3974, Nov. 2021.

[27] J. Ulbl and M. Rakusa, "Validation of Slovenian version of ADAS-Cog for patients with mild cognitive impairment and Alzheimer's disease," *Acta Neurol. Belgica*, vol. 122, no. 3, pp. 695–702, Jun. 2022.

[28] A. Bartos and S. Diondet, "The sensitive amnesia light and brief assessment (ALBA) is a valid 3-min test of 4 tasks indicative of mild cognitive deficits," *Neurología*, vol. 40, no. 6, pp. 586–598, Jul. 2025, doi: 10.1016/j.nrl.2023.02.007.
[29] H. Goodglass and E. F. Kaplan, *The Assessment of Aphasia and Related Disorders*, 2nd ed., Philadelphia, PA, USA: Lea & Febiger, 1983.
[30] M. D. Lezak, D. B. Howieson, E. D. Bigler, and D. Tranel, in *Neuropsychological Assessment* 5th ed. New York, NY, USA: Oxford University Press, 2012.
[31] A. Bartos and S. Diondet, "Sensitive written hedgehog PIC ture n aming and i mmediate r ecall (PICNIR) as a valid and brief test of semantic and short-term episodic memory for very mild cognitive impairment," *J. Alzheimer's Disease*, vol. 102, no. 2, pp. 396–410, Nov. 2024, doi: 10.1177/13872877241289385.
[32] J. Macoir and C. Hudon, "Expected performance of Quebec–French older adults on the batterie rapide de Dénomination (BARD)," *Appl. Neuropsychol., Adult*, vol. 28, no. 5, pp. 583–587, Sep. 2021.
[33] L. Zhuang, Y. Yang, and J. Gao, "Cognitive assessment tools for mild cognitive impairment screening," *J. Neurol.*, vol. 268, no. 5, pp. 1615–1622, May 2021.
[34] C. D. Patnode, L. A. Perdue, R. C. Rossom, M. C. Rushkin, N. Redmond, R. G. Thomas, and J. S. Lin, "Screening for cognitive impairment in older adults: Updated evidence report and systematic review for the U.S. preventive services task force," *JAMA*, vol. 323, no. 8, pp. 764–785, 2020, doi: 10.1001/jama.2019.22258.
[35] L. Paesen and M. Leijten, "Name agreement and naming latencies for typed picture naming in aging adults," *Clin. Linguistics Phonetics*, vol. 33, nos. 10–11, pp. 930–948, Nov. 2019, doi: 10.1080/02699206.2019.1590734.
[36] M.-C. Lin and Y.-C. Chih, "Application of the semantic fluency test in the screening of Mandarin-Chinese-speaking older adults with mild dementia of the Alzheimer type," *Behav. Sci.*, vol. 13, no. 8, p. 635, Jul. 2023, doi: 10.3390/bs13080635.
[37] G. Glosser, S. E. Kohn, R. B. Friedman, L. Sands, and P. Grugan, "Repetition of single words and nonwords in Alzheimer's disease," *Cortex*, vol. 33, no. 4, pp. 653–666, Dec. 1997, doi: 10.1016/S0010-9452(08)70723-8.
[38] C. Randolph, M. C. Tierney, E. Mohr, and T. N. Chase, "The repeatable battery for the assessment of neuropsychological status (RBANS): Preliminary clinical validity," *J. Clin. Experim. Neuropsychol.*, vol. 20, no. 3, pp. 310–319, Jun. 1998, doi: 10.1076/jcen.20.3.310.823.
[39] M. Košálová, "Diagnostika afázie v akutnÍ a subakutnÍ fázi: Mississippi aphasia screening test–Ceská verze (MASTcz)," *Neurologie Pro Praxi*, vol. 13, no. 6, pp. 314–316, 2012. [Online]. Available: https://www.neurologiepropraxi.cz/pdfs/neu/2012/06/06.pdf
[40] A. T. Beck, N. Epstein, G. Brown, and R. A. Steer, "An inventory for measuring clinical anxiety: Psychometric properties," *J. Consulting Clin. Psychol.*, vol. 56, no. 6, pp. 893–897, 1988, doi: 10.1037/0022-006x.56.6.893.
[41] J. A. Yesavage, T. L. Brink, T. L. Rose, O. Lum, V. Huang, M. Adey, and V. O. Leirer, "Development and validation of a geriatric depression screening scale: A preliminary report," *J. Psychiatric Res.*, vol. 17, no. 1, pp. 37–49, Jan. 1982, doi: 10.1016/0022-3956(82)90033-4.
[42] R. I. Pfeffer, T. T. Kurosaki, C. H. Harrah, J. M. Chance, and S. Filos, "Measurement of functional activities in older adults in the community," *J. Gerontol.*, vol. 37, no. 3, pp. 323–329, May 1982, doi: 10.1093/geronj/37.3.323.
[43] *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Washington, DC, USA: American Psychiatric Publishing, 2013, doi: 10.1176/appi.books.9780890425596.
[44] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020, *arXiv:2006.11477*.
[45] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: A free tool for segmenting, labeling and transcribing speech," in *Proc. 1st Int. Conf. Lang. Resour. Eval. (LREC)*, Granada, Spain, 1998, pp. 1373–1376.
[46] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," 2023, *arXiv:2310.11230*.
[47] J. Straková, M. Straka, and J. Hajič, "Open-source tools for morphology, lemmatization, POS tagging and named entity recognition," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Jun. 2014, pp. 13–18, doi: 10.3115/v1/p14-5003.

LUBOŠ ŠMÍDL received the M.Sc. and Ph.D. degrees from the University of West Bohemia (UWB) in Pilsen, Czech Republic, in 2000 and 2006, respectively.

He has been a member of the Department of Cybernetics, UWB, since 2000, and where he has been an Assistant Professor, since 2006. He is the Principal Investigator of the Project "Telemedicine Self-Examination of Speech and Memory for Rapid Detection of Cognitive Impairments Using Machine Learning Methods." He is working on several research and application projects. His research interests include automatic speech recognition and voice dialogue systems.

FILIP POLÁK received the B.Sc. and M.Sc. degrees in cybernetics and control engineering (artificial intelligence and automation) from the Faculty of Applied Sciences, University of West Bohemia in Pilsen, Czech Republic, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree.

He is currently a Research Assistant with the Faculty of Applied Sciences, University of West Bohemia in Pilsen. His research interests include spoken dialog systems, machine learning, and natural language understanding.

LUCIE ZAJÍCOVÁ received the M.Sc. degree in cybernetics and control engineering (artificial intelligence and automation) from the Faculty of Applied Sciences, University of West Bohemia in Pilsen, Czech Republic, in 2008, and the Ph.D. degree in cybernetics from the University of West Bohemia in Pilsen, in 2017, with a focus on information retrieval in speech data.

She is currently a Researcher at the NTIS Center, University of West Bohemia in Pilsen, specializing in artificial intelligence and machine learning. Her research interests include machine learning, information retrieval, and natural language processing.

TOMÁŠ LEBEDA received the M.Sc. degree in cybernetics and control engineering (artificial intelligence and automation) from the Faculty of Applied Sciences, University of West Bohemia in Pilsen, Czech Republic, in 2024, where he is currently pursuing the Ph.D. degree.

He is currently a Research Assistant with the Faculty of Applied Sciences, University of West Bohemia in Pilsen. His research interests include text processing, information extraction, parsing, and the manipulation of structured and unstructured data for intelligent applications.

JAN ŠVEC received the M.Sc. and Ph.D. degrees from the University of West Bohemia (UWB) in Pilsen, Czech Republic, in 2007 and 2013, respectively.

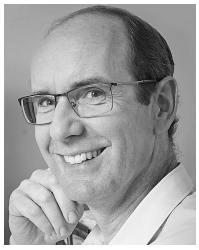
He has been a member of the Department of Cybernetics, UWB, since 2006. Since 2012, he has been a Researcher with the New Technologies for the Information Society (NTIS) Research Centre, UWB. He has contributed to projects on robust automatic speech recognition, spoken dialog systems, and archival search. His research interests include natural language processing, speech recognition, and applications of artificial neural networks.

JAN TUPÝ received the M.Sc. degree in cybernetics and control engineering (artificial intelligence and automation) from the Faculty of Applied Sciences, University of West Bohemia in Pilsen, Czech Republic, where he is currently pursuing the Ph.D. degree.

He is currently a Research Assistant with the Faculty of Applied Sciences, University of West Bohemia in Pilsen. His research interests include machine learning, large language models, and modern frontend technologies.

MARTIN BULÍN received the M.Sc. degree in robot systems from the University of Southern Denmark, Odense, Denmark, in 2016, and the M.Sc. degree in cybernetics and control engineering and the Ph.D. degree in cybernetics (network architecture search algorithms) from the University of West Bohemia in Pilsen, in 2017 and 2024, respectively.

He is currently a Researcher at the NTIS Center, University of West Bohemia in Pilsen, specializing in artificial intelligence and machine learning. His research interests include neural networks, multi-agent systems, the Internet of Things, and modern web technologies.



ALEŠ BARTOŠ received the M.D. degree (Hons.) from the First Faculty of Medicine, Charles University, Prague, Czech Republic, in 1993, and the Ph.D. degree in neuroscience from the Third Faculty of Medicine, Charles University, in 2006. He passed the first-degree advanced neurology board examination with distinction in 1996 and the second-degree examination with distinction in 1999.

He was appointed as an Associate Professor of neurology, in 2009, and a Professor of neurology, in 2019. He is a Clinical Neurologist, a Researcher, and an Academic Teacher with more than 25 years of experience in cognitive neurology and dementia. He has performed over 14 000 cognitive examinations in more than 1400 patients. He developed several innovative Czech cognitive tests, most notably the Amnesia Light and Brief Assessment (ALBA) and Picture Naming and Immediate Recall (PICNIR), both certified by the Ministry of Health of the Czech Republic, as well as magnetic resonance visual scales for early detection of neurodegenerative disorders. He founded and has led the Memory Clinic at the University Hospital Královské Vinohrady and the Third Faculty of Medicine, Charles University, since 1999. He was the Head of the Department of Cognitive Disorders, National Institute of Mental Health, from 2015 to 2021. He is the author or co-author of four monographs, 11 book chapters, and 127 Scopus-indexed publications (3840 citations and H-index of 28). His recent publications include two tests that he developed and that are used in the current grant, application, and related article: the PICNIR test for semantic and short-term episodic memory (*Journal of Alzheimer's Disease*, 2024) and the ALBA test (*Neurologia*, 2025). He is a co-investigator of the TACR Sigma Grant ''Telemedicine Self-Examination of Speech and Memory for Rapid Detection of Cognitive Impairments Using Machine Learning Methods'' (TQ01000332). He also participates in other projects, including the Charles University Neuroscience COOPERATIO Program (207038) and development grants from the Ministry of Health of the Czech Republic (University Hospital Královské Vinohrady, 00064173). His research interests include the early diagnosis of Alzheimer's disease, the development and validation of cognitive tests and questionnaires, psychometry, clinical neurochemistry, neuroimmunology, and brain imaging.

Prof. Bartoš is a member of Czech Neurological Society and a Founding Member of its Section of Cognitive Neurology. His work has received multiple awards, including five Czech Neurological Society Prizes for best original paper and one Czech Psychiatric Society Prize.