Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Locally-Run, Multimodal Screening Assistant Combining Conversational NLP, Self-Report Questionnaires, and Clock-Drawing Analysis for Low-Cost Early Cognitive-Decline Pre-Screening

Angelica Dana

angelicadana74@gmail.com

Sharda University

Aryan Kr Samanta

aryankrsamanta@gmail.com

Sharda University

Method Article

Keywords: cognitive screening, dementia, multimodal machine learning, Clock Drawing Test, natural language processing, speech recognition, self-report questionnaire, Logistic Regression

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11719517/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Early detection of cognitive decline remains one of the most persistent challenges in preventive healthcare: validated instruments such as the MMSE or MoCA require a trained clinician, a scheduled appointment, and roughly thirty minutes of structured testing, a combination of requirements that keeps most people from ever being screened until symptoms are already noticeable to others. We present a multimodal, self-administered pre-screening assistant that combines three independent, freely-available signals — (1) conversational linguistic features extracted from spoken responses via offline automatic speech recognition, (2) a structured 12-item self-report questionnaire modeled on the domain structure of the AD8 and Everyday Cognition (ECog) clinical scales, and (3) automated image analysis of a hand-drawn Clock Drawing Test — into a single feature vector scored by a trained classifier. Every component runs locally on consumer hardware using free, open-source libraries and models, with no per-query API cost. We describe the design of each signal, the specific engineering failures discovered and corrected during development and testing, and a Logistic Regression classifier trained and evaluated via leave-one-out cross-validation on a manually generated synthetic dataset of 20 simulated participant records. The final model, built from 19 features spanning all three signals after an explicit check for label leakage, reaches 85% leave-one-out accuracy against an 80% majority-class baseline. We report this result together with its central limitations — the data are manually generated synthetic data rather than recordings from human participants, and training labels were derived from the system's own questionnaire and clock-drawing thresholds rather than independent clinical diagnosis — and argue that the primary contribution of this work is a demonstrated, fully-integrated, explainable multimodal pipeline rather than a validated clinical tool.

1. Introduction

Dementia and related cognitive disorders are typically identified late in their progression, often only after a family member notices repeated memory lapses, disorientation, or a decline in day-to-day functioning [1]. By the time these signs are obvious enough to prompt a clinical visit, meaningful intervention windows may already be narrowing. A large part of this delay is not medical — it is logistical. Formal cognitive assessments such as the Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA) [2] require a trained clinician, a scheduled appointment, and roughly thirty minutes of structured testing, a combination of requirements that keeps most people from ever being screened until symptoms are already apparent to others.

This work asks a narrower, more tractable question: could a short, low-friction, self-administered digital interaction — something a person could complete from their own home in under ten minutes — surface enough signal to be worth flagging for a real clinical follow-up? The goal is explicitly not to replace clinical assessment, but to lower the barrier to a first check, in the same spirit as a home blood-pressure cuff or a glucose monitor: imperfect, but useful as an early prompt.

1.1 Contributions

- A fully-implemented, locally-run, free-to-operate three-signal screening pipeline combining conversational NLP, a validated-scale-inspired questionnaire, and automated Clock Drawing Test analysis.
- A documented account of real engineering failures — in semantic recall scoring, feature aggregation, OCR thresholding, and hand-stroke detection — found during development and testing on manually generated synthetic inputs, and the specific fixes applied.
- A trained Logistic Regression classifier, evaluated via leave-one-out cross-validation against a majority-class baseline, with an explicit label-leakage check across three feature-set variants.

1.2 Scope and Non-Goals

This system is explicitly scoped as a pre-screening aid, not a diagnostic tool. Every report the system generates carries this disclaimer. The 20-record dataset used for evaluation is manually generated synthetic data; no human participants were recruited and no real human-subject data were collected, consistent with this work's origin as a student research project rather than a clinical trial; results throughout this paper should be read accordingly.

2. Related Work and Clinical Background

2.1 The AD8 and Everyday Cognition (ECog) Scales

The AD8 (Ascertain Dementia 8-item Questionnaire) [3] was developed at Washington University in St. Louis as a brief, validated screening instrument for early cognitive impairment, asking eight yes/no-style questions about changes in memory, judgment, and daily functioning. A meta-analysis [4] reports approximately 91% sensitivity and 64% specificity for detecting dementia using the AD8, and the instrument is public domain for clinical use. The Everyday Cognition (ECog) scale [5] extends beyond memory alone, covering everyday language use, planning, organization, divided attention, and visuospatial ability, designed to capture functional decline as it shows up in daily life rather than testing memory in a clinical vacuum. Our Signal 2 questionnaire borrows the domain structure of both instruments — three questions each across Memory, Language, Daily Function, and Orientation & Attention — scored on a 1–5 frequency scale rather than the AD8's original binary format, for finer-grained input to a downstream classifier.

2.2 The Clock Drawing Test

The Clock Drawing Test (CDT) is one of the oldest and most widely used cognitive screening tasks [6], prized for how much information it captures from a single instruction: draw a clock face showing a specific time, commonly ten minutes past eleven. Correctly completing this task requires intact visuospatial planning, executive function, and numerical sequencing. The Montreal Cognitive Assessment itself incorporates a Clock Drawing Test as one component of its broader battery [2],

directly supporting the rationale of combining a drawing-based test with other modalities rather than relying on it alone.

2.3 Conversational and Linguistic Markers

A growing body of research on ‘digital phenotyping’ has examined whether naturalistic speech and language patterns — vocabulary diversity, pause frequency, filler-word usage, sentence complexity, and delayed recall accuracy — correlate with early cognitive changes, often before those changes are detectable through self-report alone [7]. Our Signal 1 draws on this research direction, treating an ordinary conversation as a source of measurable linguistic features rather than requiring a separate, formal test.

2.4 Positioning of this Work

Each of the three approaches surveyed above has documented blind spots when used alone: a questionnaire can be answered inconsistently with day-to-day function; a clock drawing can be affected by fine motor impairment unrelated to cognition; conversational features can be affected by anxiety or unfamiliarity with the assessor. Combining independent signals is a standard strategy in screening design specifically because the blind spots of one signal are unlikely to line up with the blind spots of another. The contribution of this work is not a new clinical instrument, but an engineering integration of three independently-validated screening concepts into one free, locally-run, fused pipeline — together with a transparent account of where that integration currently falls short of a clinically defensible system.

3. System Architecture

The system is organized as a set of independent Python modules, one per signal, joined by a single orchestration script that runs all three in sequence for one participant record (in this study, one simulated participant) and merges their output into a single combined row. This modular structure allowed parallel development of different signals without shared-file conflicts, and allows any one signal's internal logic to be changed later without touching the other two.

3.1 Data Flow

- The participant speaks answers to a short set of conversational prompts (Signal 1), recorded and transcribed via an offline automatic speech recognition model.
- The participant completes a 12-question structured questionnaire (Signal 2).
- The participant draws a clock, photographs it, and supplies the image (Signal 3).
- Each signal module returns a flat dictionary of named numerical/boolean features.
- An orchestration layer merges all three dictionaries (prefixed s1_/s2_/s3_ to avoid key collisions) into one row, which is appended to a growing CSV dataset and, once a classifier is trained, scored live in the same session.

3.2 Feature Summary

Table I summarizes the features produced by each signal. In total, Signal 1 contributes 10 features, Signal 2 contributes 4 domain-level features used in the final model (12 raw item-level features are also available but excluded from the final model, see Section 5.3), and Signal 3 contributes 6 features, including the two features later identified as label-construction inputs and excluded from the final evaluation.

TABLE I. FEATURES PRODUCED BY EACH SIGNAL

Signal 1 (Conversational NLP)	Signal 2 (Questionnaire)	Signal 3 (Clock Drawing)
recall_score, fluency_unique_count, fluency_repeats, filler_ratio, type_token_ratio, noun_pronoun_ratio, avg_sentence_length, recall_response_time, normal_response_time, usable_response_count	domain_Memory, domain_Language, domain_Daily_Function, domain_Orientation_Attention (used in final model); 12 raw items M1–M3, L1–L3, D1–D3, O1–O3 (used only in reduced-leak variant)	circle_present, numbers_present, numbers_complete, numbers_count_found, hands_present, hands_detected_count

4. Methodology: Signal Design and Engineering

4.1 Signal 1 — Conversational NLP

Signal 1 tells the participant a small set of unrelated facts, asks an unrelated baseline question, then asks the participant to recall the facts from memory, and finally asks a category-fluency task (naming as many animals as possible). Delayed recall is scored via semantic similarity rather than exact wording: the recall response is split into sentence-level chunks, each chunk and each target fact are embedded using a sentence-transformer model (all-MiniLM-L6-v2) [8], and each fact is matched against its single best-scoring chunk (cosine similarity threshold 0.55). An earlier implementation embedded the entire multi-fact response as one vector; this diluted the combined embedding across all mentioned topics and produced misleadingly low recall scores for genuinely correct answers — the per-chunk approach was adopted specifically to correct this failure mode, discovered during testing with a synthetic test response in which all three facts were correctly recalled in a single run.

Additional Signal 1 features include verbal fluency (unique items and repeats in the category-fluency task, with a whitespace-splitting fallback to handle comma-less spoken transcriptions), filler-word ratio, type-token ratio (vocabulary richness), noun-to-pronoun ratio (via part-of-speech tagging, using spaCy's en_core_web_sm model [9]), average sentence length, and response timing for both the baseline and recall questions.

A second engineering issue was found during testing of the full session pipeline: the session object recalculated linguistic features fresh after every turn, overwriting rather than accumulating them, so a single empty transcription (caused by an unrelated input-capture issue) silently zeroed out valid statistics from other turns. This was corrected by aggregating every non-empty response into a running list across the whole session and computing all linguistic features once, at the end, over the combined text — with empty transcriptions excluded entirely and a usable_response_count field added for transparency.

4.2 Signal 2 — Self-Report Questionnaire

Signal 2 is a 12-item structured self-report instrument spanning four domains — Memory, Language, Daily Function, and Orientation & Attention — three items each, scored on a 1–5 frequency scale. Each domain's average score is compared against thresholds adapted from AD8 cutoff conventions (1.0–2.0 normal, 2.1–3.0 mild concern, 3.1 + significant concern), and an overall score is computed as the mean across all 12 items. A doctor-facing report is generated showing each domain's score and a plain-language recommendation scaled to how many domains were flagged.

4.3 Signal 3 — Clock Drawing Test Image Analysis

Signal 3 analyzes a photograph of a hand-drawn clock for three structural elements: a circular face (via OpenCV's Hough Circle Transform), the numbers 1 through 12 (via Tesseract OCR), and two hands pointing in distinct directions, combined into a 0–3 structure score.

The first working version measured crop regions and minimum hand length as a fraction of the raw photo's pixel dimensions. Testing against a manually generated test image exposed a flaw: phone photos typically include substantial surrounding blank page, so a correctly-drawn clock photographed from further away could fail detection purely because the drawing occupied a smaller fraction of the frame. This was corrected by first locating the clock circle, then scaling every subsequent measurement relative to that circle's own detected radius rather than to the raw photo's dimensions.

Two further corrections were made to number and hand detection. First, OCR originally used one fixed global brightness threshold to binarize the image, which failed under the uneven lighting of a manually generated test photo (a shadow gradient and faint bleed-through text); this was replaced with locally-adaptive thresholding, and OCR was run across three Tesseract page-segmentation modes with results unioned together. Second, the original hand detector used probabilistic Hough line detection, searching for one continuous line segment per hand; this failed when two hands overlapped near the pivot point, an entirely normal feature of real handwriting, consistently missing the shorter hour hand. This was replaced with a radial ink-angle histogram: every dark pixel within an inner ‘hand zone’ is binned by its angle from the detected center into five-degree buckets, and contiguous clusters of buckets with sufficient ink (accounting for 0/360-degree wraparound) are each counted as one hand. Verified against the same test image, hand detection improved from 1 to 2 hands and number detection from 0 to 3 numbers after these fixes.

4.4 Voice Integration

Signal 1 was initially built around typed responses; voice input was added both for ecological validity (cognitive screening is inherently conversational) and because typed response time measures typing speed rather than speech hesitation. Offline transcription was implemented via OpenAI's Whisper model (‘base’ size) [10], chosen over a cloud-dependent alternative because Signal 1's downstream features depend directly on transcription accuracy and because offline operation avoids reliability and rate-limiting concerns across repeated test sessions. A basic average-volume check flags near-silent recordings immediately after capture, allowing a participant to redo a failed recording within the same session rather than losing that data point.

5. Experimental Setup

5.1 Synthetic Dataset

The dataset consists of 20 manually generated synthetic records, each representing one simulated participant. No human participants were recruited and no real human-subject data were collected; ethics committee approval was therefore not applicable to this study. The records were created manually so that the full pipeline could be developed and evaluated end-to-end, consistent with this work's scope as a student research project. Because the data are synthetic and are not drawn from any clinical or general population, they cannot be used to assess how the system would perform on real people.

5.2 Label Construction and a Key Limitation

No independent clinical diagnosis was available for the 20 synthetic records at the time of training. Labels were therefore derived automatically from thresholds already built into the system's own Signal 2 and Signal 3 report logic: a participant is labeled needs_attention if their Signal 2 total score exceeds 2.0 or their Signal 3 structure score is 2 or below, and healthy otherwise. This has a direct consequence for evaluation: any model trained on the exact columns used to construct the label (s2_total_score, s3_clock_structure_score) will score artificially well, since it partially re-derives arithmetic it was already handed rather than discovering an independent pattern. We treat this as the central limitation of the present evaluation and address it directly through feature engineering (Section 5.3) rather than leaving it undisclosed in a single headline accuracy figure.

5.3 Feature Engineering and Leakage Check

Three feature-set variants were compared to isolate and quantify the leakage risk described above:

- v1 — Full feature set (34 columns), including the exact label-source columns. Retained only as a reference point illustrating how misleading unaddressed leakage can be.
- v2a — Reduced-leak (32 columns): drops the two exact label-source columns but keeps all 12 raw Signal 2 items, which remain highly redundant with one another given only 20 samples.

- v2b — Final model (19 columns): drops the label-source columns and collapses the 12 raw Signal 2 items into 4 domain-level scores, combining 10 Signal 1 features, 4 Signal 2 domain scores, and 5 Signal 3 features.

5.4 Model Selection and Evaluation Protocol

Logistic Regression was selected over Random Forest given the small sample size: with 20 records and up to 34 raw features, a Random Forest has far more parameters than the data can support and tends to produce unstable, overfit splits. Logistic Regression's coefficients also remain directly interpretable, which matters for a screening tool where explaining a flagged result is as important as producing one. All variants use class_weight='balanced' within a scikit-learn Pipeline (imputation, standardization, Logistic Regression with fixed regularization C = 0.1), evaluated via leave-one-out cross-validation given the very small sample, and compared directly against a majority-class baseline (the accuracy obtained by predicting the more frequent label for every record without using any features).

6. Results

The 20-record synthetic dataset is label-imbalanced: 16 records were labeled needs_attention and 4 healthy, giving a majority-class baseline of 80%. Table II summarizes leave-one-out accuracy for each feature-set variant.

TABLE II. LEAVE-ONE-OUT CROSS-VALIDATION RESULTS

Model Variant	Features	Validation	Accuracy
Majority-class baseline	—	—	80%
v1 — Full feature set (leaky)	34 (incl. label-source cols.)	Leave-One-Out CV	85%
v2a — Reduced-leak	32 (raw Q items retained)	Leave-One-Out CV	85%
v2b — Final model	19 (10 S1 + 4 S2 + 5 S3)	Leave-One-Out CV	**85%**

All three variants, including the final leakage-reduced model (v2b), reach 85% leave-one-out accuracy, with 100% precision and 81% recall on the needs_attention class — five percentage points above the majority-class baseline, achieved without the exact columns used to construct the label. The ten features weighted most heavily by the final model's coefficients, in order, were: Signal 2's Language domain score, Signal 3's clock number count, Signal 3's circle presence, Signal 3's hand presence, Signal 2's Memory domain score, Signal 1's noun-to-pronoun ratio, Signal 1's fluency repeat count, Signal 2's Orientation & Attention domain score, Signal 1's recall score, and Signal 1's filler ratio. The presence of features from all three signals among the top predictors indicates that the final model's decision is not dominated by any single modality.

7. Discussion and Limitations

7.1 Interpreting the 85% Figure

The reported accuracy should be read as a demonstration that the full pipeline — from raw conversational, questionnaire, and image data, through feature engineering, to a trained and cross-validated classifier — functions correctly end-to-end, and that the model modestly outperforms naive guessing even after the most direct labeling leakage is removed. It should not be read as evidence of a clinically validated screening tool. The synthetic dataset is small (n = 20) and meaningfully imbalanced (16 vs. 4), and — most importantly — the data are manually generated synthetic data rather than measurements from human participants, and the labels themselves were derived from the system's own thresholds rather than independent clinical judgment, so the 85% figure partially reflects internal consistency between Signal 2/3's scoring rules and the classifier, not solely predictive validity against ground truth.

7.2 Sources of Noise

On real human input, both automatic speech recognition (Signal 1) and OCR (Signal 3) can be expected to introduce measurement noise that varies with accent, speech clarity, handwriting neatness, and photo/audio quality — independent of a participant's actual cognitive status. A sample of this size cannot fully average out this noise, and several test sessions were affected by recording-quality issues that were caught and flagged (silence detection) but not eliminated.

7.3 Self-Report Bias

As Signal 2's own generated report notes, self-reported symptom scores are known in the clinical literature to sometimes underestimate actual impairment relative to informant-reported versions of the same questionnaire — a limitation inherited directly from the AD8 methodology Signal 2 is modeled on.

7.4 Privacy and Ethical Considerations

In real-world use, voice recordings, questionnaire answers, and clock-drawing photographs would be sensitive, health-adjacent personal data. This study used only manually generated synthetic data: no human participants were recruited, no real personal data were collected, and ethics committee approval was therefore not applicable. In its current prototype form, the system stores data locally in plaintext CSV/image files rather than any secured or encrypted system; this would need to be addressed before any deployment or any study involving human participants. Every automated judgment the system produces — per-signal flags and the final classifier prediction — is paired with an explicit statement that it is a screening aid, not a diagnosis.

8. Conclusion and Future Work

We presented a multimodal cognitive pre-screening assistant combining conversational NLP, a validated-scale-inspired questionnaire, and automated Clock Drawing Test analysis, built entirely from free, locally-run tools and evaluated end-to-end on a manually generated synthetic dataset of 20 records. Building

and testing each signal end-to-end on manually generated inputs surfaced several non-obvious engineering failures (diluted semantic recall scoring, overwritten feature aggregation, photo-scale-dependent OCR and hand detection), and resolving them meaningfully improved each signal's reliability on our test inputs. A Logistic Regression classifier trained on a leakage-checked 19-feature set reaches 85% leave-one-out accuracy against an 80% baseline, drawing on all three signals rather than any single one. Because the dataset is synthetic, this figure demonstrates that the pipeline functions end-to-end rather than its performance on real individuals.

The most important next step is replacing automated, rule-derived labels with clinician-reviewed labels on a larger, ethically-approved participant pool — potentially by administering a validated instrument such as the MoCA alongside this tool for the same participants, to sanity-check its flags against a real clinical reference point. Further future work includes extending Signal 3 to score hand-position and number-placement accuracy rather than presence alone, evaluating whether larger Whisper model sizes improve transcription accuracy enough to justify their runtime cost, adding a graphical interface in place of the current terminal-based flow, and investigating secure, encrypted storage for participant data ahead of any study involving human participants. We release this work as a transparent account of both what the pipeline currently achieves and what remains before it could be considered a validated screening instrument.

Declarations

Ethics Statement

This study used manually generated synthetic data for the development and evaluation of the proposed screening system. No human participants were recruited and no real human-subject data were collected. Ethics committee approval was therefore not applicable to this study.

References

1. World Health Organization (2023) Dementia: Key Facts. WHO Fact Sheets
2. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. J Am Geriatr Soc 53(4):695–699
3. Galvin JE, Roe CM, Powlishta KK, Coats MA, Muich SJ, Grant E, Miller JP, Storandt M, Morris JC (2005) The AD8: A brief informant interview to detect dementia. Neurology 65(4):559–564
4. Meta-analysis of AD8 sensitivity and specificity for dementia detection. PMC10497164
5. Galasko D, Bennett D, Sano M, Ernesto C, Thomas R, Grundman M, Ferris S (1997) An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. Alzheimer Disease & Associated Disorders

6. Shulman KI (2000) Clock-drawing: Is it the ideal cognitive screening test? Int J Geriatr Psychiatry 15(6):548–561
7. Fraser KC, Meltzer JA, Rudzicz F (2016) Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis 49(2):407–422
8. Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP
9. Honnibal M, Montani I (2017) spaCy 2: Natural language understanding with Bloom embeddings. convolutional neural networks and incremental parsing
10. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2022) Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356
11. Pedregosa F et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
12. Bradski G (2000) The OpenCV Library. Dr. Dobb's Journal of Software Tools
13. Smith R (2007) An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)