Automated Explainable Schizophrenia Assessment from Verbal-Fluency Audio

Rok Rajher
Jure Žabkar
rr3244@student.uni-lj.si
jure.zabkar@fri.uni-lj.si
University of Ljubljana,
Faculty of Computer and Information Science,
Ljubljana, Slovenia

Abstract

Schizophrenia is associated with cognitive impairments that are difficult to assess with traditional neuro-psychological tests. Currently, these tests are manually administered by clinical doctors and rely on subjective assessment of patient's behavior, self-reported symptoms, medical history, and mental state. Recent advances in deep learning substantially improved automatic speech recognition (ASR), and large language models (LLMs), enabling the development of computational tools that can partially automate aspects of psychiatric assessment. We present the first fully automated classification of individuals with schizophrenia based on verbal-fluency tests conducted in Slovene language. Our multi-stage pipeline involves audio preprocessing, automatic transcription using the Truebar ASR model, the extraction of meaningful verbal and non-verbal features, and learning a machine learning model. The Explainable Boosting Machine (EBM) trained on the obtained feature set achieved the best overall performance.

Keywords

schizophrenia, automatic speech recognition, large language models, verbal-fluency tasks, machine learning

1 Introduction

Schizophrenia is a chronic and severe mental disorder [8, 11] that affects how a person thinks, feels, and behaves. As a psychotic disorder it is characterized by a combination of disorganized thinking and behavior, hallucinations, and delusions [2, 14]. The symptoms have major implications on individual's social life and can lead to a lifelong care [1, 7]. Schizophrenia affects about 1% of the population worldwide [9].

Currently, there is no objective or standardized diagnostic test for schizophrenia. The most widely used diagnostic frameworks in clinical practice are the DSM-5 [2] and the ICD-11 [14]. With rapid improvements in automatic speech recognition (ASR), large language models (LLMs), and machine learning, there is rising interest in computational tools that support, augment, or partially automate aspects of psychiatric assessment.

Clinicians have long noted that schizophrenia systematically affects speech in two ways:

(1) how people speak: acoustic-prosodic markers such as pause structure, speech rate, and prosodic variability, and
(2) what they say: lexical-semantic markers such as category switching, perseverations, and vocabulary diversity.

These are best observed during verbal-fluency tasks - short, standardized, low-burden, and already used in clinical practice. Our main hypothesis is that short recordings of Slovene verbal-fluency tasks contain sufficient discriminative signal, captured by acoustic and semantic features, to separate individuals with schizophrenia from healthy controls.

In this paper, we present automated machine learning pipeline for the detection and explanation of schizophrenia, leveraging the capabilities of ASR models and state-of-the-art LLMs. The tests were conducted in the Slovene language and consisted of two one-minute subtasks: (1) a semantic fluency task, where participants were asked to list as many animal names as possible, and (2) a phonetic fluency task, where participants were instructed to generate words beginning with the letter 'L'. The approach is based on audio recordings of verbal fluency tests collected by Marinković [10]. Our results can be directly compared to those reported by Marinković [10], where the transcription and analysis of the tests were performed manually. The details of our study are extensively described in [13].

2 Methods

2.1 Participants

The dataset comprises of 126 participants: 58 individuals with a clinical diagnosis of schizophrenia (SH), and 68 healthy controls (HC). All individuals in the SH group were patients admitted to the University Psychiatric Clinic Ljubljana. All participants were adults aged 18 years or older and gave consent to being part of the experiment.

Standard demographic information was collected for each participant, including age, gender, highest level of education, academic performance (school grades), marital status, and employment status. The dataset is balanced with respect to age and gender.

For participants diagnosed with schizophrenia, additional clinical information was recorded: illness duration, number of hospitalizations, and the presence of chronic or co-occurring health conditions. The median illness duration among individuals with schizophrenia was 10 years, with a median of 4 hospitalizations.

The study was approved by the Medical Ethics Committee of the Republic of Slovenia (approval number: 0120-51/2024-2711-4). All participants received a detailed explanation of the study procedures and provided written informed consent prior to participation.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.2350

Measure	SH	HC
Total Participants	58	68
Male Distribution	29	35
Female Distribution	29	33
Median Age (years)	45	46.5
Median Primary School Grade	3	5
Median High School Grade	3	4
Median Illness Duration (years)	10	–
Median Number of Hospitalizations	4	–
Prevalent Education Level	Elem.	HS
Prevalent Marital Status	Married	Married
Prevalent Employment Status	Retired	Employed

Table 1: Demographic and clinical characteristics of the participants.

2.2 Testing procedure

Each participant completed a verbal fluency test consisting of two sub-tasks:

(1) **Phonetic fluency task:** participants were asked to produce as many Slovene words as possible beginning with the letter 'L'. Proper nouns, including names of people or places, were not allowed. The task lasted 62 seconds in total: during the first 2 seconds, the letter 'L' was displayed on the screen, followed by 60 seconds for verbal response.
(2) **Semantic fluency task:** participants were instructed to name as many animals as possible in the Slovene language. Pet names and proper nouns were not allowed. The task duration was 60 seconds.

The testing procedure was standardized: each individual was seated in front of a laptop computer. After reading the instructions for the phonetic fluency task, the participant pressed a key to begin, initiating the countdown. After completing the first task, the instructions for the second task (semantic fluency) were displayed. Again, the participant initiated the task by pressing a key when ready. This concluded the verbal fluency test.

Healthy participants were tested at the Faculty of Computer and Information Science, University of Ljubljana, while individuals with schizophrenia were assessed at the University Psychiatric Clinic Ljubljana. To ensure consistency across conditions, all recordings were conducted in quiet, isolated rooms to eliminate possible noise and distractions.

All WAV files then underwent the same audio enhancement pipeline: (i) dynamic range compression to reduce variability due to speaking loudness and microphone distance, and (ii) loudness normalization to achieve consistent perceived loudness across recordings. These steps were implemented with standard functions from `pydub` and applied identically to both sites prior to feature extraction.

2.3 Data Format

The final dataset consists of 126 WAV audio recordings, one per participant, captured using the built-in laptop microphone during the test sessions. The audio tracks are encoded in uncompressed PCM format at a sampling rate of 44.1 kHz with a single (mono) audio channel. Additionally, there are 126 corresponding CSV files containing timestamps that indicate the start and end times of each subtask. Together, these audio and timestamp files serve as the primary data sources for all subsequent audio- and speech-based analyses.

3 Preprocessing

3.1 Audio Data Preparation

The WAV recordings were initially divided into two distinct audio segments using the provided timestamp files: (1) a segment corresponding to the phonetic verbal fluency task and (2) a segment corresponding to the semantic verbal fluency task.

Both audio segments were then processed through a series of audio enhancement steps:

(1) **Dynamic range compression:** To improve audio quality and ensure uniformity, downward dynamic range compression (threshold = -20.0 dBFS, ratio = 4:1, attack time = 5 ms, release time = 50 ms) was applied to each segment. This reduces the volume gap between the quietest and loudest parts of a signal [6].
(2) **Loudness normalization:** adjusting each segment to a target level of -20 dBFS. This ensured consistent perceived loudness across all recordings, reducing variability from differences in speaker volume, room acoustics, or microphone distance.
(3) **Final output:** Finally, the two fully processed segments per participant (phonetic and semantic) were saved as separate WAV files. These files constitute the final audio data used for all subsequent analyses.

All of the described steps were implemented using standard functions provided by the `pydub` library.

3.2 Feature Engineering

After automated transcriptions have been processed we performed feature engineering. Based on clinical knowledge, we created meaningful features that serve as reliable markers for distinguishing between individuals with and without schizophrenia. Three core symptoms of schizophrenia are directly applicable to our verbal-fluency tasks: disorganized speech, disorganized behavior, and negative symptoms. The primary rationale behind our feature construction is grounded in these core symptom domains.

Audio recordings are represented in two forms: (1) as text, derived from automated ASR transcriptions, and (2) as spectrograms – a visual representation of the frequency content of the audio signal over time. We constructed two groups of features:

(1) **Verbal features:** 39 features derived from the automated text transcriptions. These features aim to quantify disorganized speech, e.g. number of phrases produced per second.
(2) **Non-verbal features:** 17 features extracted directly from the spectrograms of the audio recordings, these features target prosodic elements such as pitch and vocal control, which are key indicators of negative symptoms like blunted affect and disorganized behavior; e.g. Mean pitch, representing the speaker's average vocal pitch.

3.3 Automated Transcription

The most critical step in the preprocessing of audio recordings is the generation of automated transcriptions. These ASR-derived transcriptions serve as the primary input for nearly all subsequent stages of feature extraction and machine learning analysis. We employed the ASR model `Truebar 24.05`, a state-of-the-art

speech recognition system for the Slovene language. The model was developed by the company Vitatis in collaboration with the Laboratory for Data Technologies at the Faculty of Computer and Information Science. Using Truebar API we programmatically uploaded each audio file and in response receive the corresponding transcribed words along with their start and end timestamps.

3.4 Transcription Adjustment

The output of the ASR system consists of transcribed words along with their associated timestamps. These transcriptions may include irrelevant content such as filler words. We used the `DSPy` library—a Python framework that enables declarative programming for prompting LLMs in a modular and programmatic way in combination with `GPT-4o` model. The transcription adjustment process was divided into two sequential steps:

(1) **Transcription filtering:** The raw transcription output from the Truebar ASR model was first passed to the GPT-4o model along with a description of the verbal fluency task and its rules. The model was instructed to retain only the words it considered to be relevant without modifying the words themselves.
(2) **Transcription correction:** The filtered transcription was then forwarded to the model in a second pass. With the same task context and rules provided, the model was now asked to adjust incorrectly transcribed words to what it inferred the participant likely intended to say. A word could potentially also be a neologism, we explicitly instructed the model to apply corrections only when the intended word was judged to be clear and obvious; otherwise, the word was left unchanged. For example, a misrecognized word like 'lon' would be corrected to 'slon' (elephant), whereas unclear or ambiguous cases were preserved as-is.

3.5 Adding Semantic Meaning

After filtering and correcting the transcriptions, we tagged each word with semantic annotations relevant to the verbal fluency task. These semantic features are crucial for distinguishing between HC and SH, as they capture subtle linguistic anomalies commonly associated with schizophrenia. We used `DSPy` framework in combination with the `GPT-4o` language model to perform automated semantic tagging. The model was provided with task-specific instructions and context for each word. For each transcribed word, we extracted the following semantic tags:

- **Intrusion:** The word is semantically unrelated to the target category (e.g. non-animal word during the animal naming task). Intrusions are often more frequent in individuals with schizophrenia and reflect impaired cognitive control and semantic memory organization [5].
- **Stiltedness:** Marks whether the word appears overly formal, unusual, or unnatural in everyday speech. Stilted language is a known linguistic feature in schizophrenia and may signal underlying disruptions in pragmatic language use [12].
- **Neologism:** a newly coined or nonsensical word not found in the lexicon. Neologisms are characteristic of disorganized thought and speech, and are especially relevant in schizophrenia research [3].
- **Word description (semantic task only):** A general, page-long descriptive summary of the word. For animals, this includes common features such as appearance, habitat, and behavior—providing a semantic embedding that captures how the word is typically perceived by the general population. In the case of neologisms, the semantic meaning was still applied based on what the word could plausibly represent or mean, allowing the model to assign an approximate semantic embedding even for novel or invented terms. This feature is used only for the semantic task, where meaning-based associations between words are essential.

3.6 Data Analysis

We trained and evaluated several machine learning models using these features. To ensure robust evaluation, we applied stratified 10-fold cross-validation. Performance was assessed using accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC). The *Explainable Boosting Machine (EBM)* consistently achieved the best results when trained on the full feature set. We additionally examined the top 10 most informative features to assess model interpretability. This approach enables us to understand better which et deficits are most prominent in individuals with schizophrenia and may be useful for targeted clinical interventions.

4 Results

We observe that the obtained ML models perform similarly when using the verbal (V) and non-verbal (N) feature sets separately, achieving an average AUC of 0.83 on both datasets. In the combined feature set (VN), the average performance improves across all metrics: AUC 0.86, CA 0.76, Sens. 0.69, Spec. 0.82, PPV 0.76, and F1 0.73. The EBM trained on the combined feature set (VN) achieved the best overall performance: AUC 0.90, CA 0.82, Sens. 0.76, Spec. 0.87, PPV 0.83, and F1 0.79.

To probe whether education could drive the observed performance, we examined models trained on verbal (V) and non-verbal (N) features separately, in addition to the combined set (VN). Verbal features are more likely to reflect educational attainment (e.g., lexical diversity, category switching), whereas core acoustic markers (e.g., pause structure, longest silent pause) are less dependent on education [4]. In our 10-fold CV, V and N models performed comparably, and VN performed best. This suggests that education alone is unlikely to explain the classification.

4.1 Global interpretation

The overall feature importance (FI) across the entire dataset is used for global interpretation of the model. We calculate it as the average absolute contribution of each feature across all samples:

$$\mathrm{FI}_j = \frac{1}{n} \sum_{i=1}^{n} \left| f_j(x_{i,j}) \right|, \tag{1}$$

where n is the total number of samples, and $f_j(x_{i,j})$ is the contribution of feature j for instance i. FI measures how strongly each feature influences the model's predictions on average.

Globally most important features are: (1) `comb_pho_lev2_avg` - the Levenshtein similarity between the filtered and adjusted transcriptions, which indicates impaired speech fluency, (2) `animal_tempo_max_gap_percent` - captures the longest silent pause during the semantic task, (3) `animal_sem_cont_max_coherence_index`, `animal_sem_cont_kurt_coherence_index`, and `ltest_sem_stat_min_coherence_index` - the first two capture the word-to-word coherence, while the third captures the lowest phonetic similarity between consecutive words

during the phonetic task, (4) `comb_osmile_F0From27.5Hz_stddeNorm_avg` - the standard deviation of pitch; highlights variability in vocal pitch — a marker of prosodic irregularity often observed in individuals with schizophrenia.

4.2 Local interpretation

Each individual prediction can be explained through the positive/negative contribution of each feature. Features with positive contributions increase the log-odds in favor of the schizophrenia class, while features with negative contributions decrease the log-odds, shifting the prediction toward the healthy control class. An example for a severe schizophrenia case is shown in Fig. 1

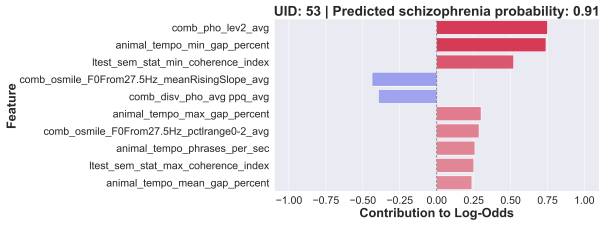

Figure 1: Local feature importance plot for a severe schizophrenia case as predicted by the EBM model. Red bars indicate contributions toward the schizophrenia class, and blue bars toward the healthy control class.

The corresponding textual explanation was generated by `GPT-4o` model: The results from the verbal fluency test indicate several features often associated with schizophrenia. Short pauses between utterances may suggest rushed or pressured speech, which can be a sign of reduced speech planning. Low semantic coherence in structured tasks may indicate the intrusion of unrelated thoughts or semantic derailment. Additionally, long pauses between utterances can reflect cognitive slowing or difficulty with word retrieval. These features collectively suggest the possibility of schizophrenia. The results suggest that, on average, the models are able to rank individuals effectively (high AUC); they can distinguish between HC and SH in terms of relative probability. The low CA, sensitivity, PPV, and F1 scores suggest that the chosen classification threshold of 0.5 may not be optimal. This issue was further addressed by evaluating the ROC curve of the best-performing model to explore whether an alternative classification threshold could improve the identification of positive cases; we observed that both the Youden-optimal threshold and the F1-optimal threshold are approximately 0.49, which differs negligibly from the used value of 0.5.

The performance of our best model, EBM, shows its strong ranking ability, and balanced classification performance on both classes.

Limitations and Future Work

Although our dataset is well-balanced, the sample size (126) is rather small; additional samples would improve model generalizability and robustness. Audio quality could be improved by using professional microphones instead of built-in laptop microphones, which would enhance transcription accuracy. Due to obtaining the audio recordings at two locations, a residual site effect cannot be fully excluded. We mitigated the risk by (i) using identical task instructions and timing in quiet rooms at both sites, (ii) applying uniform dynamic range compression and loudness normalization to all audio, and (iii) demonstrating that transcript-only models (verbal features) remain predictive, indicating that performance is not driven by background acoustics. Future studies should also include participants with other psychiatric conditions, such as major depressive disorder or bipolar disorder.

Conclusion

We developed and evaluated an automated, explainable pipeline for schizophrenia assessment using 126 verbal-fluency audio recordings (healthy controls: 68; schizophrenia: 58). The pipeline comprises audio pre-processing, automatic transcription with the Truebar ASR model, and extraction of verbal (transcript-derived) and non-verbal (acoustic/temporal) features. The features were then used in training and evaluation of several classical machine-learning models.

Across models, combining verbal and non-verbal features consistently yielded the strongest results. The Explainable Boosting Machine achieved the highest performance: CA 0.82, Sens. 0.76, Spec. 0.87, PPV 0.83, F1 0.79, and AUC 0.90. Due to the EBM's inherent interpretability, we produced global explanations and local explanations (per-instance contribution plots), complemented by GPT-4o–generated textual summaries. A high model performance and associated explanations provide a firm ground for potential decision support system in clinical practice.

5 Acknowledgments

This work was partially supported by the Slovenian Research Agency (ARIS), research program Artificial Intelligence and Intelligent Systems (Grant No. P2-0209).

References

[1] Bandar AlAqeel and Howard C. Margolese. 2012. Remission in schizophrenia: Critical and systematic review. *Harvard Review of Psychiatry* 20, 6 (2012), 281–297.
[2] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.).* American Psychiatric Publishing, Arlington, VA.
[3] Janna N. De Boer, Sanne G. Brederoo, Alban E. Voppel, and Iris E. C. Sommer. 2020. Anomalies in language as a biomarker for schizophrenia. *Current Opinion in Psychiatry* 33, 3 (2020), 212–218.
[4] J. N. De Boer, A. E. Voppel, S. G. Brederoo, H. G. Schnack, K. P. Truong, F. N. K. Wijnen, and I. E. C. Sommer. 2023. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychological Medicine* 53, 4 (March 2023), 1302–1312.
[5] Flavia Galaverna, Adrián M. Bueno, Carlos A. Morra, María Roca, and Teresa Torralva. 2016. Analysis of errors in verbal fluency tasks in patients with chronic schizophrenia. *The European Journal of Psychiatry* 30, 4 (2016), 305–320.
[6] Dimitrios Giannoulis, Michael Massberg, and Joshua D. Reiss. 2012. Digital dynamic range compressor design—A tutorial and analysis. *Journal of the Audio Engineering Society* 60, 6 (2012), 399–408.
[7] Josep Maria Haro, Diego Novick, Jordan Bertsch, Jamie Karagianis, Martin Dossenbach, and Peter B. Jones. 2011. Cross-national clinical and functional remission rates: Worldwide Schizophrenia Outpatient Health Outcomes (W-SOHO) study. *The British Journal of Psychiatry* 199, 3 (2011), 194–201.
[8] Thomas R. Insel. 2010. Rethinking schizophrenia. *Nature* 468, 7321 (2010), 187–193.
[9] Stephen R. Marder and Tyrone D. Cannon. 2019. Schizophrenia. *The New England journal of medicine* 381, 18 (2019), 1753–1761. doi:10.1056/NEJMra1808803
[10] Mila Marinković. 2024. Analysis of speech fluency in patients with schizophrenia [Master's Thesis, University of Ljubljana, Faculty of Computer and Information Science].
[11] Robert A. McCutcheon, Tiago Reis Marques, and Oliver D. Howes. 2020. Schizophrenia—An overview. *JAMA Psychiatry* 77, 2 (2020), 201–210.
[12] Victor Peralta, Manuel J. Cuesta, and Jose de Leon. 1992. Formal thought disorder in schizophrenia: A factor analytic study. *Comprehensive Psychiatry* 33, 2 (1992), 105–110.
[13] Rok Rajher. 2025. Automatic Generation of Explanations in Diagnosing Schizophrenia Using Speech Fluency Testing [Master's Thesis, University of Ljubljana, Faculty of Computer and Information Science].
[14] World Health Organization. 2022. ICD-11: 6A20 Schizophrenia. Retrieved from https://icd.who.int/browse/2025-01/mms/en#1683919430.