

# Revisiting Semantic Markers of Schizophrenia: A Machine Learning Analysis of Verbal Fluency Features

**Vaibhav Singh**

Banaras Hindu University

**Tanusree Nath**

Banaras Hindu University

**Alka Jalan**

Banaras Hindu University

**Yash Jain**

Banaras Hindu University

**Manjari Gupta**

`manjari@bhu.ac.in`

Banaras Hindu University

**Marisha .**

Banaras Hindu University

---

## Article

**Keywords:** Computational psychiatry, Machine Learning, Schizophrenia, Semantic Features, Verbal Fluency Test

**Posted Date:** August 14th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10480385/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Revisiting Semantic Markers of Schizophrenia: A Machine Learning Analysis of Verbal Fluency Features

Vaibhav Singh<sup>1</sup>, Tanusree Nath<sup>1</sup>, Alka Jalan<sup>1</sup>, Yash Jain<sup>2</sup>, Manjari Gupta<sup>1\*</sup>, Marisha<sup>1\*</sup>

<sup>1</sup> Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi 221005, Uttar Pradesh, India

<sup>2</sup> DST-Centre for Interdisciplinary Mathematical Sciences, Banaras Hindu University, Varanasi 221005, Uttar Pradesh, India

Corresponding author: Marisha (marisha@bhu.ac.in) Manjari Gupta (manjari@bhu.ac.in) (Contributed equally)

**ABSTRACT** Schizophrenia is a chronic mental illness that affects thought processes, emotional regulation, and behavior. Language impairment is one its suspected cognitive signs, including abnormalities in semantic organization and verbal fluency. Prior research has demonstrated that semantic features derived from the Verbal Fluency Test (VFT), including fluency score (FS), number of semantic switches (NSS), mean cluster size (MCS), average cluster similarity (ACS), and average switch similarity (ASS), are effective in distinguishing individuals with schizophrenia from healthy controls. In this study, we examine the relative importance of these features from a machine learning perspective. Notably, this is the first work to incorporate ACS and ASS into an artificial intelligence-driven schizophrenia detection framework. The five semantic features are first ranked according to their importance using both the Mann-Whitney U test and permutation-based feature importance scores. Subsequently, order-based feature elimination is applied to the ranked feature sets and evaluated across five machine learning classifiers. Results from both pipelines consistently indicate that ACS and ASS exhibit superior discriminative power, while MCS and NSS fail to demonstrate statistically or computationally robust importance, contradicting prevailing assumptions in the existing literature. Overall, this study provides novel insights into the significance of VFT-based semantic features for schizophrenia identification, bridging traditional clinical assessment and AI-driven diagnostic frameworks.

**KEYWORDS** Computational psychiatry; Machine Learning; Schizophrenia; Semantic Features; Verbal Fluency Test.

## 1. INTRODUCTION

Schizophrenia is a severe and persistent mental illness diagnosed in around 0.5-1.0% of the world's population over a lifetime [1]. The disorder is characterized by profound disruptions in the process of thoughts, perception, emotion, and social relations. While major symptoms such as hallucinations and delusions are widely recognized, another less visible but highly informative indicator of schizophrenia is language disorganization. This linguistic disruption reflects deeper cognitive deficits and can be measured using standardized tools like verbal fluency tests [2].

The Verbal Fluency Test (VFT) is a widely used neuropsychological assessment tool for language production that requires participants to generate words in response to a cue [3, 4]. It is commonly used to measure executive function over semantic memories. In its standard form, participants are instructed to produce as many words as possible from a particular category, for example, animals, within a limited time that is usually 60 seconds. VFT performance is often impaired in individuals with schizophrenia, who may exhibit reduced word production, frequent repetitions [5], lowered psychomotor speed [6] and disorganized semantic groupings [7]. These patterns suggest disruptions

in semantic memory access and cognitive flexibility, both of which are central to the schizophrenia pathology [8].

The traditional scoring of VFT mainly focuses on the total number of correct responses (Fluency Score), which fail to capture the underlying semantic structure and temporal dynamics of word generation. In addition to FS, semantic features derived from VFT, like Mean Cluster Size (MCS), Number of Semantic Switches (NSS), Average Cluster Similarity (ACS) and Average Switch Similarity (ASS) have also proven to have strong discriminating power in identifying diseases like schizophrenia, dementia and related forms of cognitive impairments ([2], [3], [5], [6]).

Despite their high clinical relevance, the utilization and understanding of these semantically rich features remains limited in automated diagnostic systems. Although recent artificial intelligence-based frameworks have begun to incorporate such features, they often focus on their performance in clinical diagnosis, without a principled understanding of their relative importance in a computational learning context. This study addresses this gap by systematically analyzing these features from both practical and computational perspectives, and by providing a grounded assessment of their relative importance within a machine learning framework. While features such as fluency score (FS), Mean Cluster Size (MCS), and Number of Semantic Switches (NSS) have been previously incorporated into automated frameworks for schizophrenia diagnosis [36, 37], more advanced semantic similarity-based features, including Average Cluster Similarity (ACS) and Average Switch Similarity (ASS), remain largely confined to clinical and descriptive data analyses [12, 35].

The overall contributions of this study are summarized as follows:

- i. **Dataset annotation and feature derivation:** We annotate the publicly available verbal fluency trajectories dataset [25], into schizophrenic and healthy control groups, and extract five key semantic verbal fluency features, namely FS, MCS, NSS, ACS, and ASS.
- ii. **Computational assessment of feature relevance:** We systematically evaluate the relative importance of each feature using both statistical analyses and machine learning-based methodologies, providing a comprehensive understanding of their computational significance.
- iii. **Novel use of similarity-based features in ML:** To the best of our knowledge, this is the first study to integrate ACS and ASS into a machine learning framework for schizophrenia diagnosis, and our findings demonstrate that these similarity-based features exhibit stronger discriminative capability compared to conventionally used measures such as FS, MCS, and NSS.

The rest of the paper is organized as follows: Section 2 talks about the background and literature involving verbal fluency impairment related to schizophrenia and studies in schizophrenia detection. Section 3, explains the methodology, including dataset & annotation, data preprocessing, feature extraction, feature analysis, and machine learning-based schizophrenia identification. Section 4 summarizes the results, section 5 discusses the interpretation of the results, section 6 concludes the study, and finally, section 7 outline potential future work in this domain.

## 2. BACKGROUND AND RELATED WORK

Extensive studies have analyzed the cognitive and linguistic impairments associated with schizophrenia, with a special emphasis on semantic processing and verbal fluency. They aim to understand the pattern in language-related abnormalities often associated with schizophrenia

patients. They also aim to explore how computational approaches can be used to enhance the diagnosis and prediction of the disease.

## 2.1 VERBAL FLUENCY AND SEMANTIC IMPAIRMENT IN SCHIZOPHRENIA

Semantic impairment is a well-documented feature of schizophrenia, and verbal fluency tasks have been widely used to assess disruptions in semantic memory and executive functioning [34]. A study by Paulsen et al. [13] demonstrated how language problems occur and compared how these problems differ based on the age of onset and type of schizophrenia. However, their study has certain limitations: unequal response count between groups, small sample size, and challenges in distinguishing age-influenced symptoms from schizophrenic traits.

Further investigations examined the relationship between verbal fluency performance and cognitive deficits in schizophrenia. Ojeda et al. [14] analysed the data to determine whether schizophrenic cognitive impairment is correlated to similar dynamics as healthy controls. Their findings identified association between specific cognitive domains and fluency performance; however, the cross-sectional design and use of chronically hospitalized patients limited both the generalizability and the statistical power.

Brébion et al. [15] examined gender variation in verbal fluency performance and its association with working memory and processing speed. While the study provided insights into sex-specific cognitive profiles, interpretation was limited to a small female subsample, with lack of control for educational background and unexamined medication effects.

Meta-analytic evidence further supports semantic impairment in schizophrenia. Bokatz et al. [16] reported that individuals with schizophrenia struggle more with category fluency than letter fluency, suggesting greater problems with semantic fluency. However, small sample sizes and unreliable task measures made it hard to draw clear conclusions about the underlying causes of these difficulties.

## 2.2 COMPUTATIONAL ANALYSIS OF VERBAL FLUENCY

To address limitation of traditional scoring methods, several studies have introduced computational approaches for analyzing verbal fluency data. Barattieri di San Pietro et al. [2] proposed a computational procedure based on vector space models to improve clustering and semantic switching behaviour in schizophrenia spectrum disorders. Their goal was to characterize fluency productions using algorithmic methods alongside traditional scoring approaches. However, the study of their sample mainly comprised older, hospitalized patients with a long history of illnesses, limiting the ability of the findings to broader patient populations. On top of that, manual labelling proved time-consuming, reducing feasibility for clinical implementation.

In recent years, more automation and data-driven linguistic analysis methods have been introduced. Bambini et al. [17] applied a semi-automated multidimensional linguistic analysis combined with clustering techniques to examine clinical heterogeneity in schizophrenia. Their results indicated that different linguistic profiles were associated with symptom severity and functional outcomes. However, limited sample diversity, and the absence of progressive validation restrict the stability and generalizability of identified profiles over time.

## 2.3 SEMANTIC SIMILARITY AND EMBEDDING-BASED APPROACHES

Advancement in computational linguistics have enabled quantitative assessment of semantic similarity [33]. Ku et al. [18], applied semantic similarity measures to verbal fluency data in first-episode psychosis, demonstrating that the patients were exhibiting derailment and tangentiality. Despite its contribution, the study relied on specific assessment instruments,

which did not serve as a standard for disorder assessment. Thus, significant intercorrelations among output measures were reported, limiting interpretability.

Holshausen et al. [19] investigated the contribution of semantic coherence in disorganised thought processes among older patients with schizophrenia by utilising latent semantic analysis to improve assessment sensitivity. Their findings linked reduced semantic coherence to formal thought disorder, supporting the utility of latent semantic measures. However, the focus on older patient population limits the applicability to younger or acutely ill populations.

Mayr et al. [6] criticized the way verbal fluency tasks are scored, arguing that common methods can't separate problems with clustering and switching from slower thinking speed, thereby reducing diagnostic specificity.

Mandera et al. [20] tested prediction-based semantic similarity models for psycholinguistic research, while also facing challenges related to data availability and the use of subjective similarity ratings.

Elvevåg et al. [21] applied latent semantic analysis to measure incoherence in patient speech, showing that semantic similarity between sentence pairs could capture language disorganization in schizophrenia. This work supports the validity of semantic vector-based analysis as a diagnostic aid, including the Word2Vec-based feature extraction.

## 2.4 PREDICTIVE AND NLP APPROACHES

Beyond descriptive analysis, machine learning and automated NLP tools have demonstrated strong predictive potential in psychiatric research [37, 38]. Bedi et al. [22] showed that semantic coherence and related linguistic features extracted from free speech could predict psychosis onset in high-risk individuals, months in advance. This highlights the predictive value of disorganized speech with the use of cosine similarity and clustering for early-stage detection [11].

More generally, Calvo et al. [23], explored the transferability of word embedding models such as, Word2Vec and GloVe from general-purpose corpora to mental health contexts, emphasizing semantic shifts and challenges in clinical interpretation. More recently, Fusaroli et al. [24] introduced graph neural network-based representations for modelling word sequences and transitions in verbal fluency tasks, demonstrating improved group discrimination relative to traditional latent semantic approaches. Despite their promise, these models require large annotated datasets, which limits generalizability in smaller clinical settings.

## 3. METHODOLOGY

This section describes the overall methodology including dataset annotation, preprocessing, feature extraction, feature analysis and model training.

### 3.1 DATASET AND ANNOTATION

The dataset used in this study was obtained from a publicly available dataset. This included 31 individuals diagnosed with schizophrenia (SZ) from NHS community psychosis clinics in London and 29 healthy control participants from the same area. However, the labeling for SZ and healthy controls were not specified in the dataset. Diagnoses were confirmed using the Structured Clinical Interview for DSM-IV-TR Axis I Disorders (SCID-I). The two groups were matched for age, gender, IQ, and education level. After applying exclusion criteria, such as age over 45, use of anticonvulsants or benzodiazepines, poor vision, lack of English education, and other psychiatric or neurological conditions, a total of 53 participants (26 with SZ and 27 healthy controls) completed the verbal fluency tasks. These tasks were conducted during the first study visit, along with cognitive and clinical assessments. Only the category fluency task data were

used in the present analysis. The dataset consists of the following columns: ‘participants’ ‘I’d’, ‘verbal responses’ and ‘target disease label’ and each row represents the verbal fluency test data for each participant.

Owing to the fact that dataset labels were missing in the original dataset, it was annotated by three independent clinical psychologists, based on their expertise. This was accomplished by having each annotator create their own independent set of annotations.

Agreement between the annotators was evaluated using Fleiss' Kappa ( $\kappa$ ) inter-annotator metric [39]. The computed  $\kappa$  value was calculated to be 0.88, indicating good agreement according to established interpretation criteria, and confirming the reliability of the annotation process [39].

The final ground-truth labels assigned to the dataset was determined by using a majority voting method amongst the annotators. This method reduces the impact of individual annotator’s biasness and results in a robust, high-quality labeled dataset used for subsequent experiments.

### 3.2 PREPROCESSING

A pre-trained 300-dimensional Word2Vec model (Google News Corpus) was loaded via GenSim to convert words and phrases into dense semantic vector representations. These representations were then preprocessed by converting them into lowercase and trimming whitespace. Additionally, replacement strategy was applied to compound terms, such as, “gold raven” that are absent from the Word2Vec vocabulary. These terms are replaced with simpler terms like “raven”, which are recognized by the vocabulary. Words that are still out of vocabulary post-replacement were discarded. The preprocessed and cleaned list was then used to derive VFT-based feature values, that are used for further experiments. Word2Vec has been widely demonstrated to effectively capture semantic similarity between words [9, 10] and with machine learning classifiers as well [27], and its representations are based on the distributional hypothesis and negative-sampling training strategy [26].

### 3.3 FEATURE EXTRACTION

To characterize the semantic significance of VFT-responses of participants, five computational features were derived following preprocessing. These were chosen because of their established validity through multiple research studies reporting their ability to detect individuals diagnosed with schizophrenia (SZ) and cognition-related deficits [28]. These features captured aspects of both traditional productivity-oriented metrics as well as cognitive-oriented metrics that provide a more complete view of the relationship between semantic organization and thoughts occurring during VFT.

#### 1) Fluency Score (FS)

It is defined as the number of unique valid words produced by each participant [29]. This reflects lexical productivity and is a common baseline metric in verbal fluency tasks, as shown in **Eq. (1)**.

$$FS = |UniqueWords| \quad (1)$$

where, UniqueWords is the set of all unique and valid responses.

#### 2) Number of Semantic Switches (NSS)

The count of local minima in the cosine similarity sequence, each indicating a transition between semantically distinct word clusters [30]. This feature reflects cognitive flexibility and the ability to shift between subcategories as shown in **Eq. (2) -(4)**.

Heuristic Rule: Switch at index  $j$  if:

$$\text{sim}(w_{j-1}, w_j) > \text{sim}(w_j, w_{j+1}) < \text{sim}(w_{j+1}, w_{j+2}) \quad (2)$$

$$NSS = \sum_{j=2}^{n-2} [\text{sim}_{j-1} > \text{sim}_j < \text{sim}_{j+1}] \quad (3)$$

where,  $\text{sim}_j = 1 - |\cos(\vec{w}_j, \vec{w}_{j+1})|$  (4)

### 3) Mean Cluster Size (MCS)

The average number of words grouped between semantic switches. Larger clusters may indicate more efficient exploitation of semantic neighborhoods, while smaller clusters suggest frequent switching [31] as shown in **Eq. (5)**.

$$MCS = \frac{1}{k} \sum_{i=1}^k |C_i| \quad (5)$$

where,  $C_i$  is the  $i^{th}$  semantic cluster formed between consecutive semantic switches cluster and  $k$  is the total number of clusters.

### 4) Average Cluster Similarity (ACS)

The mean cosine similarity of adjacent word pairs within each cluster. High values suggest tight semantic cohesion and effective use of related vocabulary [32] as shown in **Eq. (6)**.

$$ACS = \frac{1}{N_c} \sum_{(i,j) \in \text{ClusterPairs}} \cos(\vec{w}_i, \vec{w}_j) \quad (6)$$

where, ClusterPairs includes all adjacent word pairs within the same cluster and  $N_c$  is the number of such pairs.

### 5) Average Switch Similarity (ASS)

The average cosine similarity of adjacent word pairs that span a detected semantic switch. Lower values typically reflect more pronounced shifts in topic or subcategory, aligning with divergent semantic exploration as shown in **Eq. (7)**.

$$ASS = \frac{1}{N_s} \sum_{(i,j) \in \text{SwitchPairs}} \cos(\vec{w}_i, \vec{w}_j) \quad (7)$$

where, SwitchPairs includes all word pairs that span identified switches, and  $N_s$  is their count.

## 3.4 STATISTICAL FEATURE ANALYSIS USING MANN-WHITNEY U TEST

A Mann-Whitney U statistical test was used to assess importance of individual VFT-based features, distinguishing schizophrenia from healthy control. Mann-Whitney U test is a non-parametric statistical procedure that is well-suited to the study of cognitive and behavioural measures, because, unlike other parametric tests, such as t-tests, the Mann-Whitney U test does not assume normal distributions of the analysed features, is robust to outliers, and can also accommodate small sample sizes.

#### 1) Mathematical Formulation

Let  $x = \{x_1, x_2, x_3, \dots, x_n\}$  and  $y = \{y_1, y_2, y_3, \dots, y_n\}$  represent the feature values for the schizophrenia and healthy control groups, respectively.

The Mann–Whitney U statistic for group x is computed as shown in **Eq. (8)**:

$$U_x = R_x - \frac{n_1(n_1 + 1)}{2} \quad (8)$$

where,  $R_x$  denotes the sum of ranks assigned to observations in group x, and  $n_1$  is the sample size of group x. The corresponding statistic for group y,  $U_y$ , is computed analogously, and the smaller of  $U_x$  and  $U_y$  is used for significance testing.

Under the null hypothesis, the U statistic follows a sampling distribution, enabling computation of a p-value.

#### 2) Hypothesis Definition:

For all VFT-based features, the following hypotheses were evaluated:

- Null hypothesis ( $H_0$ ): The distribution of the feature values in patients with schizophrenia and healthy controls is identical.
- Alternative hypothesis ( $H_1$ ): The distribution of the feature values in patients with schizophrenia and healthy controls is significantly different.

The lower the p-value, the stronger the evidence against the null hypothesis, and therefore, the greater likelihood that the feature is discriminative between these two groups.

| Features                    | U     | p-value | Effect Size (r) |
|-----------------------------|-------|---------|-----------------|
| Fluency Score               | 243.0 | 0.0554  | 0.264           |
| Number of Semantic Switches | 263.0 | 0.1168  | 0.215           |
| Mean Cluster Size           | 319.5 | 0.5812  | 0.077           |
| Average Cluster Similarity  | 210.5 | 0.0127  | 0.343           |
| Average Switch Similarity   | 245.0 | 0.0604  | 0.259           |

TABLE 1. Mann–Whitney U Test results for VFT features

**Table 1** shows the results of the Mann–Whitney U test applied to individual VFT features to examine whether their distributions differ significantly between patients with schizophrenia and healthy controls. The table reports the U statistic and the p-value for each feature. Among the features, Average Cluster Similarity (ACS) shows a statistically significant group difference ( $p = 0.0127$ ), indicating strong evidence against the null hypothesis and suggesting that this feature is highly discriminative between the two groups.

While other features, like Fluency Score (FS) and Average Switch Similarity (ASS), demonstrate trend-level differences with p-values close to the conventional significance threshold ( $p \approx 0.05$ ), implying moderate discriminative potential. In contrast, Number of Semantic Switches (NSS) and Mean Cluster Size (MCS) yield higher p-values, indicating weaker or non-significant differences between schizophrenia and healthy control groups.

### 3) Effect Size Estimation:

Beyond statistical significance, the magnitude of group differences was calculated using the effect size  $r$ , computed as shown in **Eq. (9)**:

$$r = \frac{Z}{\sqrt{n}} \quad (9)$$

where,  $Z$  is the standardized test statistic derived from the U value, and  $N = n_1 + n_2$  is the total number of observations across both groups.

The effect size  $r$  quantifies how well the predictor features separate out the outcome features. It had the following levels of interpretation:

- Small ( $r \approx 0.1$ )
- Medium ( $r \approx 0.3$ )
- Large ( $r \geq 0.5$ )

### 4) Feature Ranking:

This pattern confirms that features with larger effect sizes show stronger group differences: MCS has the smallest effect size, reflecting limited sensitivity to group-level language disorganization; NSS, ASS, and FS show trend-level, medium effect sizes; and ACS shows a significant effect size, confirming that this embedding-based semantic measure captures meaningful disruption in semantic organization. These effect sizes therefore indicate which features are most worth retaining for the subsequent classification analysis.

### 5) Experimental Design:

A machine-learning pipeline was implemented to evaluate the feature importance revealed by the Mann-Whitney U Test. Five machine learning classifiers: Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Gaussian Naïve Bayes (GNB) and Extreme Gradient Boosting (XGB) were used to evaluate the features and determine their discriminating power in identifying schizophrenic patients. For each

classifier, the experiment was conducted iteratively using a decremental feature selection strategy, where features were progressively removed based on their lower effect sizes as estimated by the Mann-Whitney U test. For example, in the first iteration, all features were retained; in the second iteration, the feature exhibiting the smallest effect size was removed.

For the experiments, the dataset was split into an 80:20 ratio, with 80% of the data used for training and the remaining 20% for testing. Macro-F1 score was used as the primary evaluation metric.

### 3.5 MACHINE LEARNING MODEL-SPECIFIC FEATURE ANALYSIS USING PERMUTATION IMPORTANCE SCORE

This section discusses permutation importance score-based VFT-feature analysis, evaluated using the same five machine learning classifiers introduced in Section 3.4(5).

#### 1) Mathematical Formulation

In addition to the Mann-Whitney U test, permutation feature importance was computed for each classifier described in the previous subsection to provide complementary, classifier-specific insights into feature relevance. To assign the permutation-based importance to each feature, firstly a baseline accuracy is calculated from retaining all the features in the training data. Next, the values of a single feature were randomly shuffled across all samples while keeping the remaining features intact. The resulting classification accuracy is compared with the baseline accuracy and their difference is noted as the corresponding feature's importance. If the difference is large, then the corresponding feature is considered to have significant power in discriminating schizophrenia from healthy controls. If the difference is near zero, then the feature does not contribute much in classification performance. A negative difference indicates noise. The feature importance values were summarized using their mean and standard deviation across multiple iterations. It can be seen across the classifiers in **Table 2**.

| Model | Feature         | ACS      | ASS      | FS       | NSS      | MCS      |
|-------|-----------------|----------|----------|----------|----------|----------|
| RF    | Importance_Mean | 0.251818 | 0.141651 | 0.020536 | 0        | 0.073939 |
|       | Importance_STD  | 0.102037 | 0.083578 | 0.069332 | 0        | 0.036970 |
| SVM   | Importance_Mean | 0.356211 | 0.313991 | 0.199860 | 0.018485 | 0.067778 |
|       | Importance_STD  | 0.111718 | 0.096929 | 0.071715 | 0.036970 | 0.040872 |
| kNN   | Importance_Mean | 0.299725 | 0.168931 | 0.032727 | 0.036177 | 0.017677 |
|       | Importance_STD  | 0.109392 | 0.103952 | 0.059573 | 0.077042 | 0.048867 |
| GNB   | Importance_Mean | 0.075035 | 0.065476 | 0.021417 | 0.060188 | 0.081761 |
|       | Importance_STD  | 0.102288 | 0.080192 | 0.105636 | 0.108448 | 0.108176 |
| XGB   | Importance_Mean | 0.124796 | 0.121114 | 0.016197 | 0        | 0.009347 |
|       | Importance_STD  | 0.093667 | 0.056717 | 0.072536 | 0        | 0.074142 |

TABLE 2. Permutation feature importance statistics (mean and SD) across classifiers

**Figure 1** represents the permutation importance scores plotted on a heatmap across classifiers and features. Brighter color show higher mean importance values, indicating greater sensitivity to classification performance.

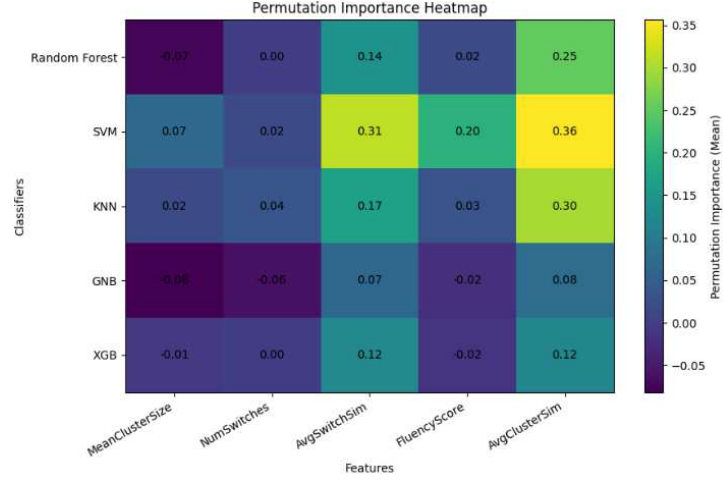


FIGURE 1. Permutation importance heatmap of VFT-based features across classifiers

Across all classifiers, Average Cluster Similarity (ACS) consistently exhibited highest importance values, as reflected by brighter regions in the heatmap. This indicates that permuting this feature leads to substantial performance degradation, further strengthening the fact that this feature being the most important. Average Switch Similarity (ASS) and Fluency Score (FS) closely follow ACS, as visible through the heatmap. On the other hand, features such as Mean Cluster Size (MCS) and Number of Switches (NSS) show near-zero or negative importance for several classifiers, suggesting limited and redundant contribution.

## 2) Experimental Design

Similar to Section C, a decremental feature selection strategy was implemented based on classifier-specific permutation feature importance scores. Only the mean values were considered to order the features and evaluate them across five machine learning classifiers that are mentioned in Section C.

## 3.6 MODEL TRAINING AND HYPERPARAMETER SETTINGS

Both the machine-learning pipelines discussed in previous subsections C and D were implemented using the same five classifiers introduced in Section 3.4(5). A brief overview of each model is discussed below:

### 1) Random Forest (RF)

Random Forest is an ensemble learning model that constructs multiple trees and aggregate their predictions to improve generalization and reduce overfitting. In this study, the Random Forest classifier consisted of  $T = 200$  decision trees. The final prediction for an input feature vector  $x$  is obtained using majority voting:

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^T, \quad (10)$$

where  $h_t(\cdot)$  denotes the prediction of the  $t^{th}$  decision tree.

Each decision tree is grown to a maximum depth of  $D_{\max} = 5$ . At each node, the optimal split is selected by minimizing the Gini impurity, defined as:

$$G = 1 - \sum_{c=1}^C p_c^2, \quad (11)$$

where  $p_c$  is the proportion of samples belonging to class  $c$  at that node.

To address class imbalance, class weights are incorporated during training. The weighted Gini impurity is computed as:

$$G_w = 1 - \sum_{c=1}^C w_c p_c^2, \quad (12)$$

where  $w_c$  denotes the weight assigned to class  $c$ , inversely proportional to its class frequency.

A fixed random seed (random\_state=42) is used to ensure reproducibility of the bootstrap sampling and feature selection processes.

## 2) Support Vector Machine (SVM)

Support Vector Machine is a margin-based classifier that uses a hyperplane for segregating between classes. The primary optimization problem is defined as:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (13)$$

subject to:

$$y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad (14)$$

where  $w$  is the weight vector,  $b$  is the bias term,  $\xi_i$  are slack variables allowing misclassification, and  $C$  is the regularization parameter controlling the trade-off between margin maximization and classification error.

To handle non-linearly separable data, a sigmoid kernel function is employed, defined as:

$$K(x_i, x_j) = \tanh(\gamma x_i^\top x_j + r), \quad (15)$$

where  $\gamma$  controls the curvature of the decision boundary and  $r$  is the kernel bias term. In this study, the parameters were set to  $C = 1.0$  and  $\gamma = 0.01$ .

### 3) k-Nearest Neighbor (kNN)

k-Nearest Neighbor is a non-parametric classifier, that assigns classifies based on majority vote of the neighboring samples. In this work, the kNN classifier was implemented with k set to 5 nearest neighbors.

The Euclidean distance between a test sample  $x$  and a training sample  $x_i$  is computed as:

$$d(x, x_i) = \sqrt{\sum_{j=1}^d (x_j - x_{ij})^2}. \quad (16)$$

Let  $N_5(x)$  denote the set of the five nearest neighbors of  $x$ . A distance-weighted voting scheme is applied, where closer neighbors exert higher influence. The weight assigned to each neighbor is defined as:

$$w_i = \frac{1}{d(x, x_i) + \epsilon}, \quad (17)$$

here  $\epsilon$  is a small constant added to avoid division by zero.

The predicted class label is obtained as:

$$\hat{y} = \arg \max_{c \in \{0,1\}} \sum_{i \in N_5(x)} w_i I(y_i = c), \quad (18)$$

where  $I(\cdot)$  is the indicator function.

### 4) Gaussian Naïve Bayes (GNB)

Gaussian Naïve Bayes is a probabilistic classifier that applies Bayes' theorem assuming conditional independence amongst the features given the class label. The posterior probability of a class  $y \in \{0,1\}$  given a feature vector  $x = (x_1, x_2, \dots, x_d)$  is expressed as:

$$P(y | x) = \frac{P(y) \prod_{j=1}^d P(x_j | y)}{P(x)}. \quad (19)$$

Under the conditional independence assumption, each feature likelihood  $P(x_j | y)$  is modelled using a Gaussian distribution:

$$P(x_j | y = c) = \frac{1}{\sqrt{2\pi\sigma_{jc}^2}} \exp\left(-\frac{(x_j - \mu_{jc})^2}{2\sigma_{jc}^2}\right), \quad (20)$$

where  $\mu_{jc}$  and  $\sigma_{jc}^2$  denote the mean and variance of feature  $x_j$  estimated from the training samples of class  $c$ .

$$\hat{y} = \arg \max_{c \in \{0,1\}} P(y = c) \prod_{j=1}^d P(x_j | y = c). \quad (21)$$

Prior to model training, all features were standardized to zero mean and unit variance to ensure numerical stability and improve classification performance:

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}, \quad (22)$$

where  $\mu_j$  and  $\sigma_j$  represent the mean and standard deviation of feature  $x_j$  computed from the training data.

##### 5) Extreme Gradient Boosting (XGB)

XGBoost is a classifier that constructs an ensemble of decision trees in a sequential manner, optimizing classification performance through gradient-based loss minimization. In this study, the XGBoost classifier was implemented with 200 estimators, a maximum depth of 3, and a learning rate of 0.1.

The model prediction for the  $i^{\text{th}}$  sample is given by:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F, \quad (23)$$

where,  $f_k$  denotes the  $k^{\text{th}}$  regression tree and  $F$  is the space of all possible trees.

For binary classification, the logistic objective is used:

$$p_i = \sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}, \quad (24)$$

where  $p_i$  is the predicted probability of the positive (schizophrenia) class.

The loss function is defined as the binary logistic loss:

$$\ell(y_i, p_i) = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]. \quad (25)$$

The overall objective function minimized during training is:

$$L = \sum_{i=1}^N \ell(y_i, p_i) + \sum_{k=1}^K \Omega(f_k) \quad (26)$$

where, the regularization term is given by:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2, \quad (27)$$

Here,  $T$  is the number of leaf nodes in a tree and  $w_j$  is the weight of the  $j^{\text{th}}$  leaf. To handle class imbalance, a scale-positive weighting factor is applied to the loss of the positive class:

$$\ell'(y_i, p_i) = \begin{cases} \alpha \ell(y_i, p_i), & \text{if } y_i = 1, \\ \ell(y_i, p_i), & \text{if } y_i = 0, \end{cases} \quad (28)$$

where  $\alpha$  denotes the scale positive weight.

During training, stochastic subsampling is applied to reduce overfitting:

$$x_i \sim D_{\text{subsample}}, |D_{\text{subsample}}| = 0.8N, \quad (29)$$

and column subsampling selects a random subset of features:

$$x_i^{(s)} \subset x_i, |x_i^{(s)}| = 0.8d. \quad (30)$$

By embedding both statistical and model-specific feature rankings directly into the experimental workflow, the study systematically evaluated how classification performance evolved as feature dimensionality was reduced under two distinct relevance criteria. This design enabled direct comparison between population-level statistical discriminability and classifier-specific feature utilization, providing complementary insight into the contribution of semantic fluency features to schizophrenia classification.

#### 4. RESULTS

This section presents the results obtained from the model training pipeline specified in section 3.6.

##### 4.1 EXPERIMENT 1

Following the Mann-Whitney U test-based ranking and order-based feature elimination strategy described in Section 3.4(5), the performance of the five classifiers under progressively reduced feature subsets is shown in **Figure 2**.

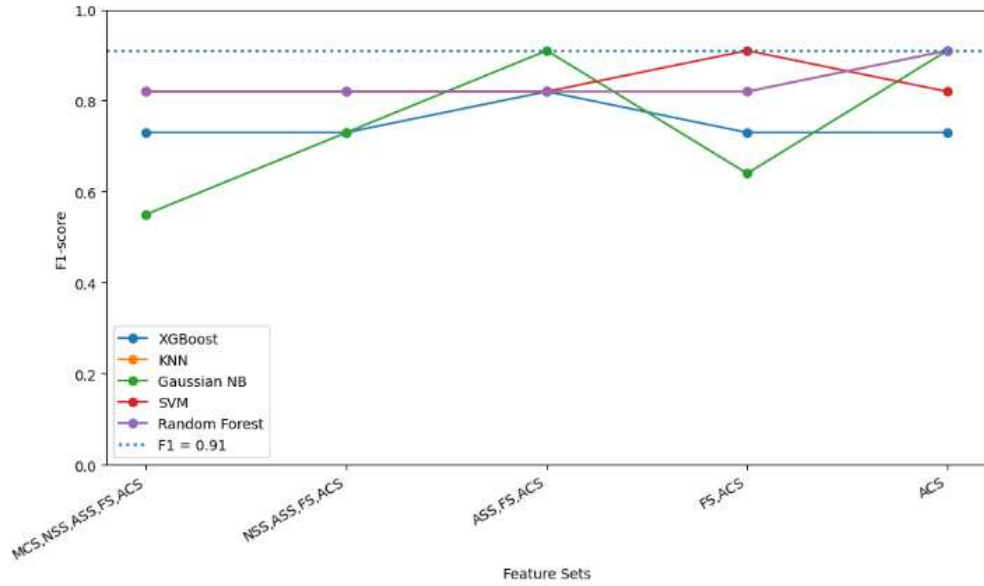


**FIGURE 2.** F1-score comparison across classifiers with decreasing feature sets based on Mann Whitney U Test

It is observed from the figure that, as the feature dimensionality decreased, classifiers retained stable performance with higher-ranked features. This indicates that the classification accuracy

was most affected and controlled by a high-quality small feature-set rather than the full feature set. Notably, the classifiers maintained competitive F1-scores (**see Figure 2**) even when only one or two top-ranked features (ACS and ASS) were used, demonstrating robustness and superiority over other features (MCS, NSS and FS) that were used in prior studies. RF, GNB and KNN (overlapped with RF) showed that using ACS alone attained the highest accuracy.

## 4.2 EXPERIMENT 2

Following the permutation-importance-based ranking and elimination strategy described in Section 3.5(2), this experiment evaluates how progressive feature reduction affects classification performance for each classifier.

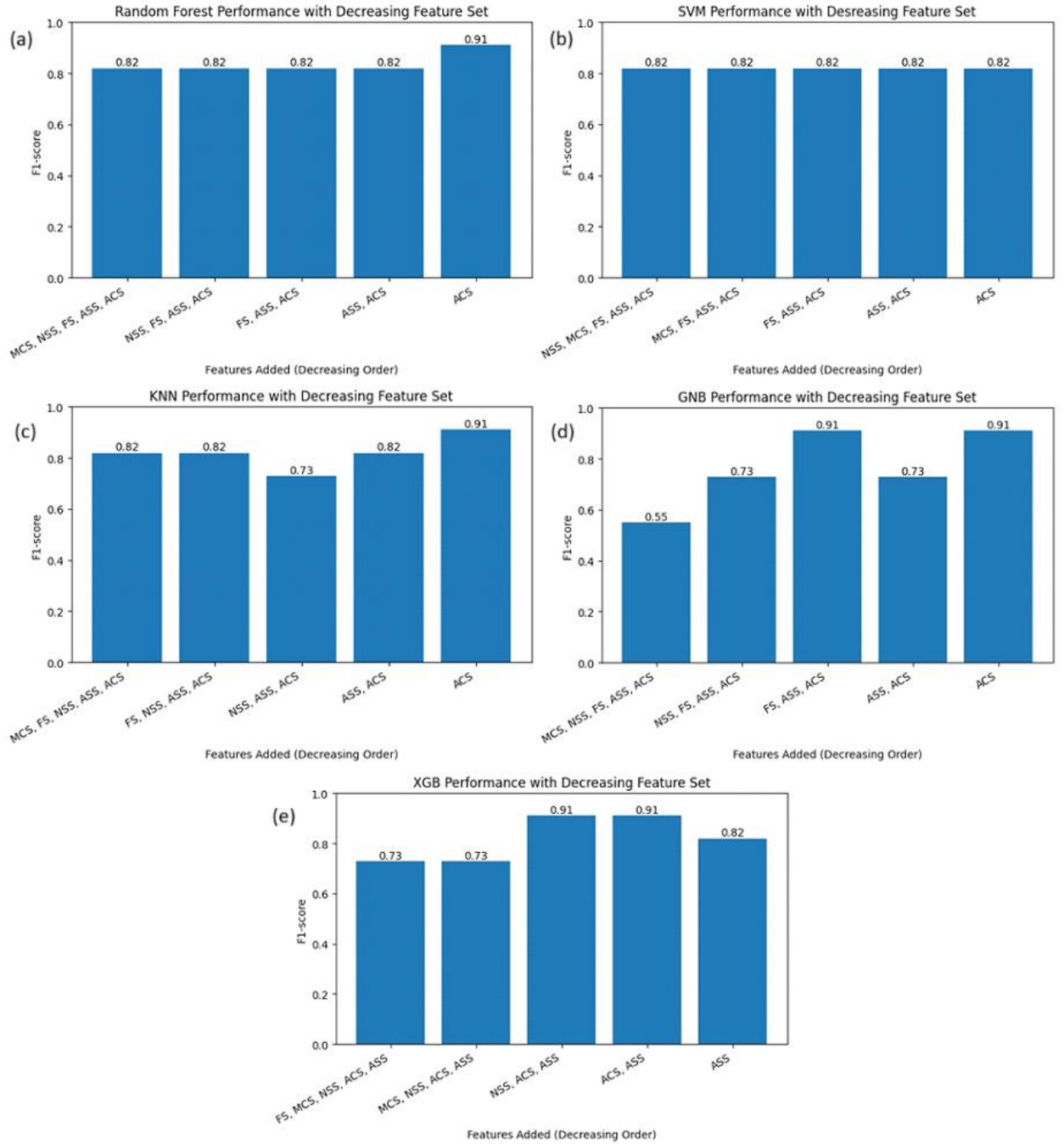


**FIGURE 3.** F1-score comparison across classifiers with decreasing feature sets based on permutation importance scores.

Overall, a consistent pattern emerged across classifiers: classification performance did not significantly degrade with decreasing feature dimensionality. In several cases, removing lower-ranked features led to improved F1 scores (**see Figure 3**), indicating the presence of feature redundancy and noise, and highlighting the importance of compact feature representations.

For the RF classifier, F1-score remained relatively stable ( $\sim 0.82$ ) when multiple features were included, demonstrating robustness to noisy inputs. However, an improved F1-score ( $\sim 0.91$ ) was observed only when ACS was retained, demonstrating its ability as a singular discriminative signal. It also aligns with the heatmap results obtained from the statistical feature assessment, where ACS consistently displays the highest importance.

In the case of SVM, it is quite evident that all the features provide similar F1-score ( $\sim 0.82$ ), being unchanged across all feature subset. This flat response to feature elimination indicates that the decision boundary was already well-defined with early feature combinations, and that SVM was relatively insensitive to feature selection in this setting.

The KNN classifier, exhibited strong sensitivity to feature dimensionality. Performance lowered substantially to ( $\sim 0.73$ ) when lower-ranked features were included and improves sharply and peaked to ( $\sim 0.91$ ) when the feature set was reduced to ACS only. This indicates the curse of dimensionality inherent to distance-based method which is sensitive to noisy or redundant features and highlights stronger reliance on dominant semantic representations.

For GNB, classification performance shows greater variability across feature subsets. Lower performance of ( $\sim 0.55$ ) was observed when multiple low-importance features are included, whereas peak F1-scores ( $\sim 0.91$ ) are achieved when the feature set is restricted to top-ranked semantic similarity features, particularly ACS alone, or in combination with FS and ASS. This behavior reflects the independence assumptions of GNB and its susceptibility to less informative features.

Finally, in XGB, F1-score remains comparatively robust across reduced feature sets, with an optimal result of ( $\sim 0.91$ ), achieved when retaining the highest-ranking semantic features, particularly combinations of ASS and ACS. Although XGB benefits from multiple features, its performance with minimal feature subsets indicates effective exploitation of semantic similarity information.

Across all classifiers, Average Cluster Similarity (ACS) consistently emerges as the most discriminative feature. Feature subsets containing ACS alone or in combination with Average Switch Similarity (ASS) frequently matched or outperformed larger feature sets, often achieving peak F1-scores ( $\approx 0.91$ ). While, features such as Mean Cluster Size (MCS), Number of Semantic Switches (NSS), and Fluency Score (FS) contributed limited or even negative marginal gains, reinforcing the dominant role of embedding-based semantic similarity measures.

This result demonstrates that classifier performance does not necessarily increase with the number of features. Instead, a compact, carefully selected feature subset: dominated by ACS, consistently yields superior and comparable performance across diverse learning paradigms, while also reducing overfitting risk, improving generalization, and lowering computational complexity.

## 5. DISCUSSION

The comparative analysis between the two experiments provides insightful information on the importance of verbal fluency-based semantic features. The Mann-Whitney U test approach highlights statistical separation amongst the features, based on their distributions. It yields an independent ranking of the features that can be clinically interpreted. In contrast, permutation

importance highlights how the features contribute to predictive performance with respect to a specific classifier, based on learned decision boundaries.

Both pipelines demonstrate that the most significant semantic similarity feature for the classification of schizophrenia is Average Cluster Similarity (ACS), followed by Average Switch Similarity (ASS). ACS alone has the potential to deliver excellent classification accuracy on RF, kNN and GNB models. This suggests that the use of semantic similarity features such as the ACS and ASS are reliable independent indicators of the presence of disorganized language in people diagnosed with schizophrenia.

In summary, the result from the analyses indicates that embedding-based semantic features dominate the models' classification's ability, whereas the traditional methods of calculating fluency and semantic switch counts have only minor effects.

## **6. CONCLUSION**

The study investigates the relevant importance of Verbal Fluency Test (VFT)-based features for machine learning based schizophrenia detection. The overall pipeline includes dataset annotation, preprocessing of verbal responses, extraction of both traditional and semantic similarity-based VFT features, followed by feature-importance assessment using both statistical Mann-Whitney U test and machine learning-based permutation importance. Both of these feature importance assessments were evaluated using five machine learning models through order-based feature elimination, enabling a robust and comprehensive analysis of feature discriminability.

Across both the experiments, Average Cluster Similarity (ACS) consistently emerged as the most significant and discriminative feature, followed by Average Switch Similarity (ASS). Notably, reducing the feature set to ACS alone, or to the combined ACS + ASS, did not degrade classifier performance. In several cases, it led to improvements, with ACS often yielding the highest F1-scores as a single-feature input. This coherence between statistical feature importance and classifier-level importance highlights the robustness and reliability of semantic similarity-based measures in capturing language disorganization associated with schizophrenia, extending their established clinical use into an automated diagnostic setting.

Moreover, traditionally used VFT measures such as Fluency Score (FS), Mean Cluster Size (MCS), and Number of Semantic Switches (NSS), demonstrated limited or inconsistent contribution across both the experiments. Despite their widespread use in prior literature, these features neither exhibited strong statistical separation nor meaningful impact on the classifier performance. This contradictory insight challenges prevailing assumptions regarding their diagnostic relevance and focuses on the need to re-evaluate commonly used VFT metrics through computationally grounded validation rather than convention alone.

Overall, this study provides novel insights into the semantic understanding of verbal fluency impairment in schizophrenia, demonstrating that semantic features such as ACS and ASS offer superior discriminative power compared to traditional measures. By merging clinical interpretability with machine learning, the proposed framework advances the understanding of VFT-based semantic markers for future AI-driven neuropsychiatric diagnostic systems.

## **7. FUTURE WORK**

Future studies should consider incorporating other domain-specific word embeddings, such as GloVe and ELMo, to enhance semantic representation. Expanding the dataset to include larger and more diverse participant samples would strengthen the model's statistical power and

generalizability. Including Thought, Language, and Communication (TLC) scale and Clinical Language Assessment of Neuropsychiatric Disorders (CLANG) alongside semantic fluency could provide a more comprehensive linguistic profile of participants. This approach may also be extended to other neuropsychiatric conditions such as bipolar disorder, major depressive disorder and PTSD, enabling comparisons between shared and disorder-specific semantic patterns. Such analyses could support differential diagnosis and early detection. Finally, integrating multimodal data, such as acoustic features from task recordings and neurocognitive data, may further enhance predictive performance and provide a more comprehensive evaluation approach.

## DATA AVAILABILITY STATEMENT

The dataset used for this study, is available in `matthewnour/verbal_fluency_trajectories` and can be accessed via this link [https://github.com/matthewnour/verbal\\_fluency\\_trajectories](https://github.com/matthewnour/verbal_fluency_trajectories). While the annotated data used to support the findings of this study and the code used can be made available upon the reasonable request from the corresponding author.

## REFERENCES

- [1] J. McGrath, S. Saha, J. Welham, et al., “A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology,” *BMC Med.*, vol. 2, p. 13, 2004, doi: 10.1186/1741-7015-2-13.
- [2] C. Barattieri di San Pietro, C. Luzzatti, E. Ferrari, G. de Girolamo, and M. Marelli, “Automated clustering and switching algorithms applied to semantic verbal fluency data in schizophrenia spectrum disorders,” *Lang. Cogn. Neurosci.*, vol. 38, no. 7, pp. 950–965, 2023, doi: 10.1080/23273798.2023.2178662.
- [3] Q. Zhao, Q. Guo, and Z. Hong, “Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment,” *Neurosci. Bull.*, vol. 29, pp. 75–82, 2013.
- [4] M. Van Beilen, M. Pijnenborg, E. H. Van Zomeren, R. J. Van Den Bosch, F. K. Withaar, and A. Bouma, “What is measured by verbal fluency tests in schizophrenia?,” *Schizophr. Res.*, vol. 69, no. 2–3, pp. 267–276, Aug. 2004, doi: 10.1016/j.schres.2003.09.007.
- [5] J. D. Henry, J. R. Crawford, and L. H. Phillips, “Verbal fluency performance in dementia of the Alzheimer’s type: a meta-analysis,” *Neuropsychologia*, vol. 42, no. 9, pp. 1212–1222, 2004, doi: 10.1016/j.neuropsychologia.2004.02.001.
- [6] U. Mayr, “On the dissociation between clustering and switching in verbal fluency,” *Neuropsychologia*, vol. 40, no. 5, pp. 562–566, Jan. 2002, doi: 10.1016/S0028-3932(01)00132-4.
- [7] E. Tyburski, A. Sokołowski, M. Chęć, J. Pełka-Wysiecka, and A. Samochowiec, “Neuropsychological characteristics of verbal and non-verbal fluency in schizophrenia patients,” *Arch. Psychiatr. Nurs.*, vol. 29, no. 1, pp. 33–38, 2015, doi: 10.1016/j.apnu.2014.09.009.
- [8] M. M. Nour, D. C. McNamee, Y. Liu, and R. J. Dolan, “Supporting Information for Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts,” [Online]. Available: <https://terpconnect.umd.edu/~toh/spectrum/PeakFindingandMeasurement.html>.

- [9] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2vec model analysis for semantic similarities in English words," *Procedia Comput. Sci.*, vol. 157, pp. 160–167, 2019.
- [10] S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis," *Expert Syst. Appl.*, vol. 152, p. 113401, 2020.
- [11] F. Günther, C. Dudschig, and B. Kaup, "Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies," *Q. J. Exp. Psychol.*, vol. 69, no. 4, pp. 626–653, 2016.
- [12] M. M. Nour, D. C. McNamee, Y. Liu, and R. J. Dolan, "Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 42, 2023, doi: 10.1073/pnas.2305290120.
- [13] J. S. Paulsen et al., "Impairment of the semantic network in schizophrenia," *Psychiatry Res.*, vol. 63, no. 2–3, pp. 109–121, Jul. 1996, doi: 10.1016/0165-1781(96)02901-0.
- [14] N. Ojeda et al., "Verbal fluency in schizophrenia: Does cognitive performance reflect the same underlying mechanisms in patients and healthy controls?," *J. Nerv. Ment. Dis.*, vol. 198, no. 4, pp. 286–291, Apr. 2010, doi: 10.1097/NMD.0b013e3181d61748.
- [15] G. Brébion et al., "Verbal fluency in male and female schizophrenia patients: Different patterns of association with processing speed, working memory span, and clinical symptoms," *Neuropsychology*, vol. 32, no. 1, pp. 65–76, Jan. 2018, doi: 10.1037/neu0000394.
- [16] C. E. Bokatz and T. E. Goldberg, "Letter and category fluency in schizophrenic patients: a meta-analysis," *Schizophr. Res.*, vol. 64, no. 1, pp. 73–78, Nov. 2003, doi: 10.1016/S0920-9964(02)00282-7.
- [17] V. Bambini et al., "Deconstructing heterogeneity in schizophrenia through language: a semi-automated linguistic analysis and data-driven clustering approach," *Schizophrenia*, vol. 8, no. 1, p. 102, Nov. 2022, doi: 10.1038/s41537-022-00306-z.
- [18] B. S. Ku, L. Pauselli, M. A. Covington, and M. T. Compton, "Computational linguistic analysis applied to a semantic fluency task: A replication among first-episode psychosis patients," *Psychiatry Res.*, vol. 304, p. 114105, Oct. 2021, doi: 10.1016/j.psychres.2021.114105.
- [19] K. Holshausen, P. D. Harvey, B. Ellevåg, P. W. Foltz, and C. R. Bowie, "Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia," *Cortex*, vol. 55, pp. 88–96, Jun. 2014, doi: 10.1016/j.cortex.2013.02.006.
- [20] P. Mandera, E. Keuleers, and M. Brysbaert, "Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation," *J. Mem. Lang.*, vol. 92, pp. 57–78, Feb. 2017, doi: 10.1016/j.jml.2016.04.001.
- [21] B. Ellevåg, P. W. Foltz, D. R. Weinberger, and T. E. Goldberg, "Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia," *Schizophr. Res.*, vol. 93, pp. 304–316, 2007, doi: 10.1016/j.schres.2007.03.001.
- [22] G. Bedi et al., "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophr.*, vol. 1, p. 15030, 2015, doi: 10.1038/npjpsych.2015.30.

- [23] R. A. Calvo and D. N. Milne, "Natural language processing in mental health applications using non-clinical texts," *Nat. Lang. Eng.*, vol. 23, no. 5, pp. 649–685, 2016.
- [24] R. Fusaroli et al., "Graph neural representations of verbal fluency tasks reveal markers of schizophrenia," *J. Neural Eng.*, vol. 19, p. 056002, 2022.
- [25] GitHub - matthewnour/verbal\_fluency\_trajectories, Accessed: Jul. 16, 2025. [Online]. Available: [https://github.com/matthewnour/verbal\\_fluency\\_trajectories](https://github.com/matthewnour/verbal_fluency_trajectories).
- [26] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.
- [27] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," in *Proc. IEEE ICCICC\**, Beijing, China, 2015, pp. 136–140, doi: 10.1109/ICCI-CC.2015.7259377.
- [28] A. K. Troyer, M. Moscovitch, and G. Winocur, "Clustering and switching as two components of verbal fluency," *Neuropsychology*, vol. 11, no. 1, pp. 138–146, 1997, doi: 10.1037//0894-4105.11.1.138.
- [29] J. H. Cerhan et al., "Diagnostic utility of letter fluency, category fluency, and fluency difference scores in Alzheimer's disease," *Clin. Neuropsychol.*, vol. 16, no. 1, pp. 35–42, 2002.
- [30] R. E. Von Studnitz and D. W. Green, "The cost of switching language in a semantic categorization task," *Biling. Lang. Cogn.*, vol. 5, no. 3, pp. 241–251, 2002.
- [31] K. Kaneko, "Clustering, coding, switching, hierarchical ordering, and control in a network of chaotic elements," *Phys. D: Nonlinear Phenom.*, vol. 41, no. 2, pp. 137–172, 1990.
- [32] M. Jakobsson and N. A. Rosenberg, "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure," *Bioinformatics*, vol. 23, no. 14, pp. 1801–1806, 2007.
- [33] T. T. Hills, M. N. Jones, and P. M. Todd, "Optimal foraging in semantic memory," *Psychol. Rev.*, vol. 119, no. 2, pp. 431–440, 2012.
- [34] N. Raoux et al., "Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study," *Cortex*, vol. 44, no. 9, pp. 1188–1196, 2008.
- [35] Lundin, N. B., Jones, M. N., Myers, E. J., Breier, A., & Minor, K. S. (2022). Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis. *Psychiatry research*, 309, 114404.
- [36] König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., & Robert, P. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and geriatric cognitive disorders*, 45(3-4), 198-209.
- [37] Gómez-Valadés, A., Martínez, R., & Rincón, M. (2024). Designing an effective semantic fluency test for early MCI diagnosis with machine learning. *Computers in Biology and Medicine*, 180, 108955.

[38] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

#### **ETHICS APPROVAL & CONSENT TO PARTICIPATE**

The present study is an analysis of a publicly available, fully anonymized dataset. Ethical approval and informed consent were obtained by the original data collectors in accordance with institutional and national guidelines. No additional ethical approval was required for the current analysis.

#### **AUTHOR CONTRIBUTIONS STATEMENT**

**Vaibhav Singh:** Data curation, Formal analysis, Visualization, Validation, Methodology, Writing – original draft. **Tanusree Nath:** Data curation, Formal analysis, Visualization, Validation, Methodology, Writing – original draft. **Alka Jalan:** Conceptualization, Methodology. **Yash Jain:** Conceptualization, Methodology. **Marisha:** Supervision, Project administration, Writing – review and editing. **Manjari Gupta:** Supervision, Project administration, Writing – review and editing.

#### **ADDITIONAL INFORMATION**

The Authors have declared that there are no conflicts of interest in relation to the subject of this study.

#### **FUNDING**

No funding was received for this study.