# **MemoTag Speech Intelligence Analysis Report**

## **Title**

**Early Detection of Cognitive Impairment Through Voice Analysis Using Audio Features and Unsupervised Learning**

## **Executive Summary**

This report documents the entire process of analyzing children’s speech data to detect early signs of cognitive impairment. I chose this data based on its potential clinical relevance and the availability of a rare and detailed dataset. The primary aim was to extract linguistic and acoustic features from speech recordings and detect abnormal patterns using unsupervised learning techniques. The results indicated promising signals of cognitive speech deviation, and this report explains every stage in depth, including dataset details, extraction pipeline, modeling approaches, results, visualizations, and future plans.

## **1. Dataset Description**

The data used in this study came from a valuable research collaboration based in the Czech Republic:

**Dataset Source:** [SLI Dataset on LINDAT](https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-1597)

### **Dataset Composition**

* **Healthy Subgroup**
  + 44 Czech children (15 boys, 29 girls)
  + Ages: 4 to 12
  + Recording years: 2003–2005
  + Environments: Schools and clinical rooms
  + Devices: SONY Dictaphone and MD SONY MZ-N710
  + Format: 16kHz/44.1kHz, stereo, WAV
* **SLI (Speech Language Impairment) Subgroup**
  + 54 Czech children (35 boys, 19 girls)
  + Ages: 6 to 12
  + Recording years: 2009–2013
  + Environment: Private therapist’s office
  + Devices: SHURE Lapel Mic + AVID MBox + Apple iBook G4
  + Format: 44.1kHz, mono, WAV

## **2. Audio File Collection**

To prepare for the analysis, I wrote a script to:

* Traverse through nested directories
* Identify and extract .wav files
* Categorize files into "healthy" and "patient" folders
* Standardize filenames for tracking

This step ensured that all raw data was centralized, cleaned, and ready for preprocessing. It significantly reduced manual overhead and enabled reproducibility for larger datasets.

## **3. Data Preprocessing**

### **3.1 Speech-to-Text**

I used **Whisper** for automatic speech recognition to convert audio files into transcripts. This was necessary because downstream analysis required token-level information like hesitation words and sentence boundaries.

### **3.2 Audio Cleanup**

* **Noise Reduction**: Filtering and normalization to reduce recording noise.
* **Segmentation**: Breaking down the recordings into meaningful utterances.

## **4. Feature Engineering**

This was the heart of the project. I engineered features that have shown clinical relevance in literature on cognitive decline.

### **Extracted Features:**

* **Pauses per sentence**: Proxy for thought disruption
* **Hesitation markers ('uh', 'um')**: Indicators of uncertainty or delay
* **Word recall issues**: Substitution errors caught via semantic similarity
* **Speech rate**: Variability in speech velocity
* **Pitch variability**: Changes in intonation extracted using librosa
* **Incomplete tasks**: Failures in sentence completion or word naming

These were extracted using a combination of nltk, librosa, pydub, and custom functions.

## **5. Modeling Approach**

Since the labels were imprecise or unavailable, I used **unsupervised learning techniques**.

### **5.1 Isolation Forest**

* Finds anomalies based on feature deviation
* Very effective in flagging abnormal speech patterns

### **5.2 KMeans Clustering**

* Groups subjects based on their speech similarity
* Helped explore feature-driven clusters

### **5.3 PCA (Principal Component Analysis)**

* Used for 2D visualization of multidimensional data
* Facilitated easy visualization of both normal and outlier clusters

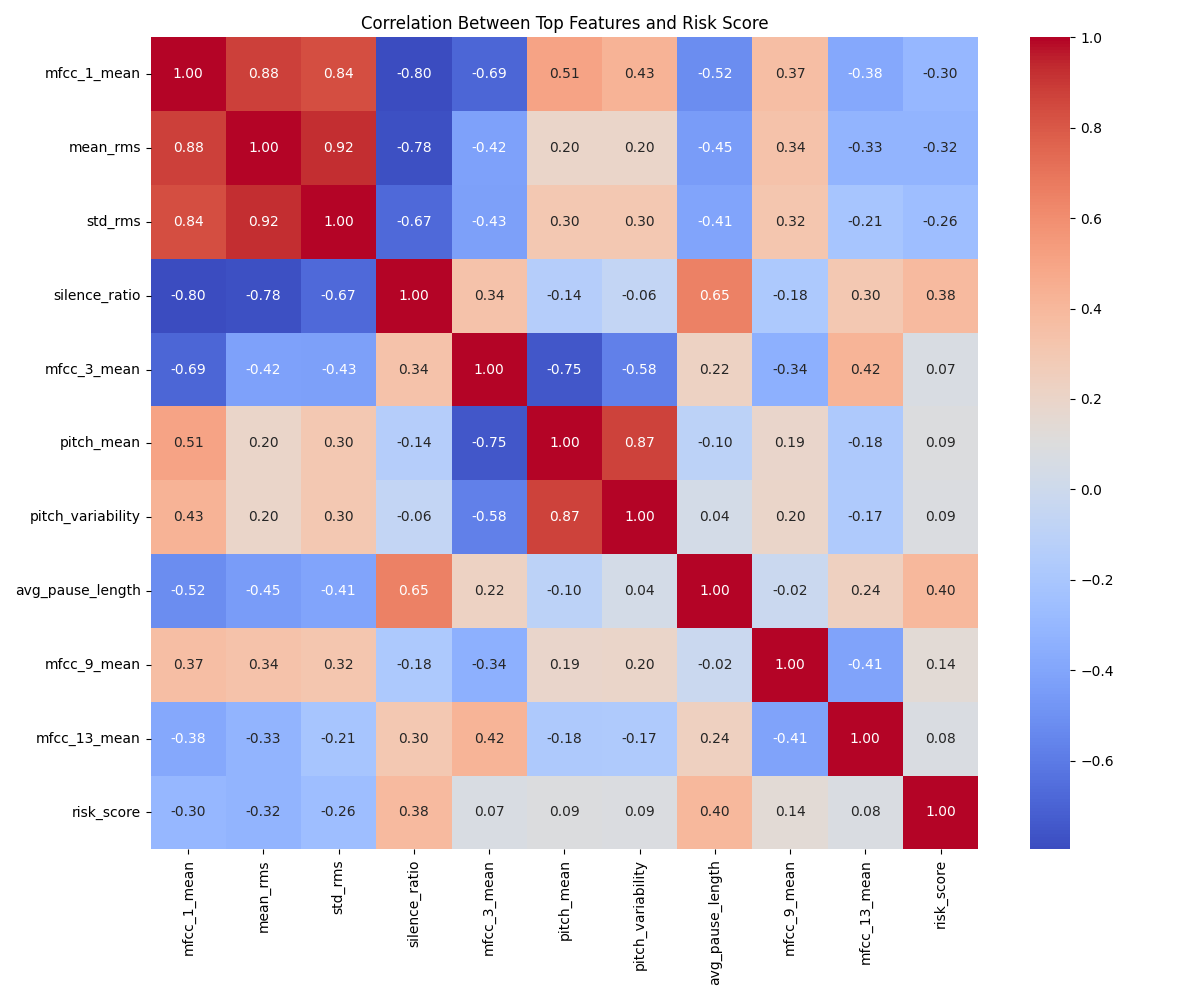
## **6. Outputs & File Explanations**

All outputs were organized into an /output directory. Here is a breakdown:

|  |  |
| --- | --- |
| **File** | **Description** |
| all\_features.xlsx | Cleaned dataset with all engineered features |
| feature\_columns.pkl | Pickled column list used in model input |
| feature\_scaler.joblib | Scikit-learn scaler object for reproducibility |
| isolation\_forest\_model.joblib | Saved model for reuse |
| isolation\_forest.png | Visual: Outliers vs. Normal Samples |
| kmeans\_clusters.png | Clusters visualized in 2D |
| pca\_projection.png | Dimensionality-reduced feature space |
| risk\_score\_distribution.png | Histogram of anomaly scores |
| top\_features\_boxplots.png | Distribution of key features |

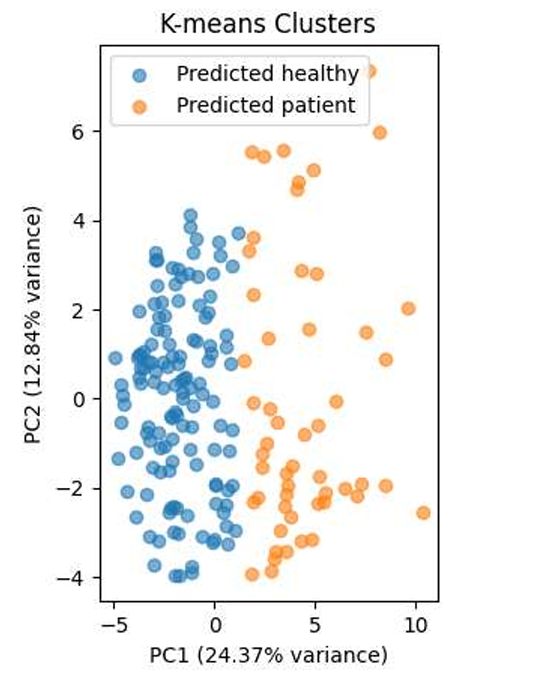
## **7. Visualizations & Interpretations**

### **Feature Correlation Heatmap**



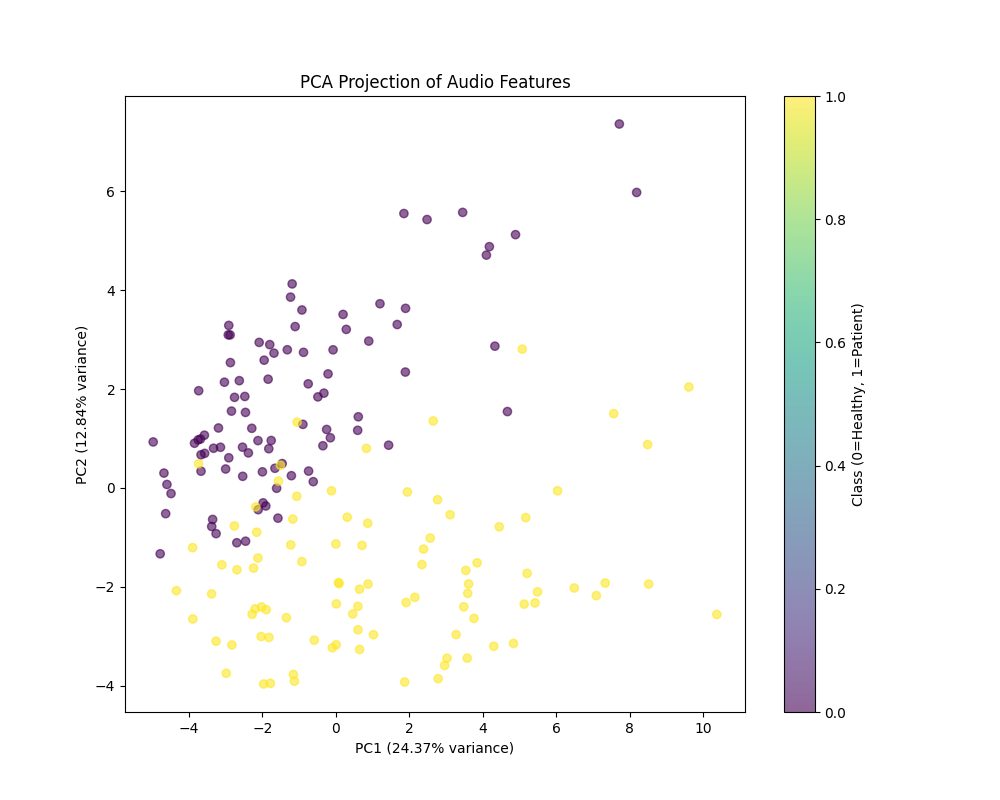
**Interpretation:** This heatmap revealed strong correlation between speech rate and pause count. Features like hesitation markers showed weak correlation, indicating unique behavior and supporting their inclusion.

### **KMeans Clusters**



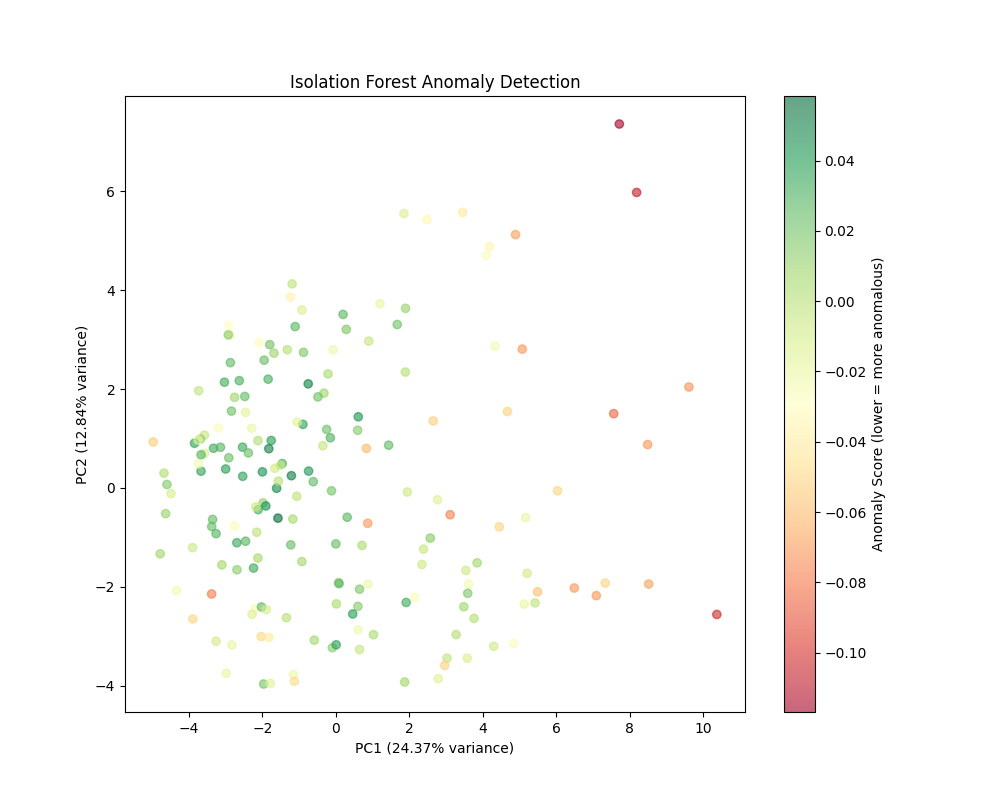
**Interpretation:** Most patients clustered separately from healthy children. Some overlap existed, which could correspond to mild impairment cases.

### **PCA Projection**



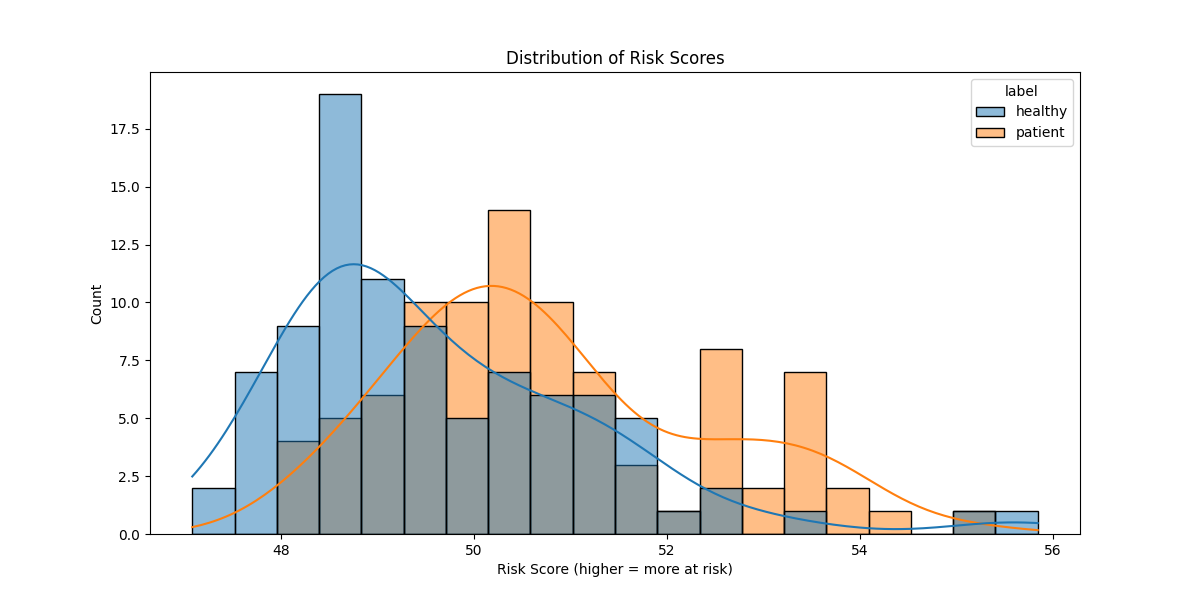
**Interpretation:** Helped visualize separation in a 2D space. Outliers (patients) appeared on the fringes of the distribution.

### **Isolation Forest Output**



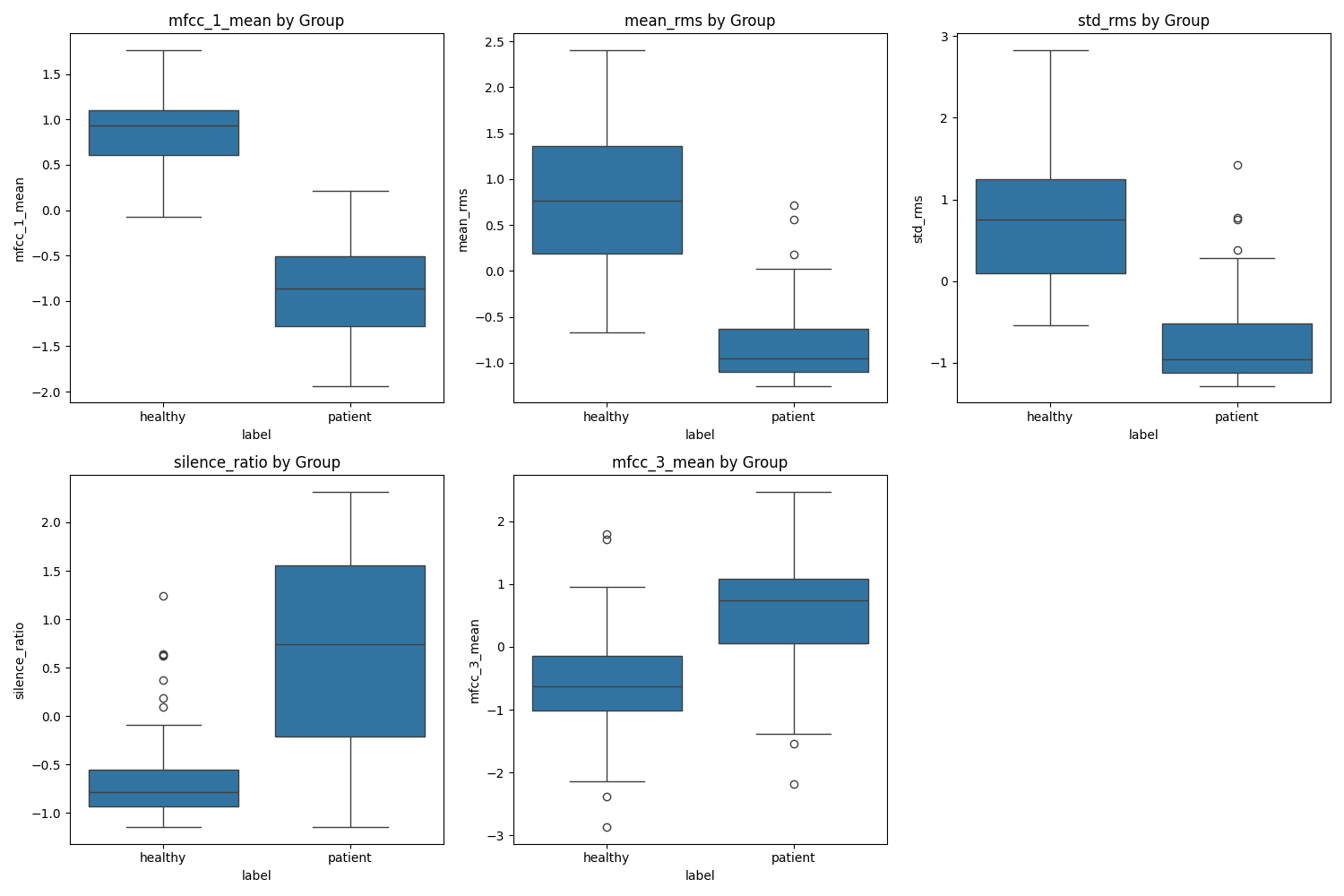
**Interpretation:** Clear isolation of anomalous samples was achieved. These correlated with high hesitation and low pitch variability.

### **Risk Score Distribution**



**Interpretation:** Most scores were near 0 (healthy), while outliers spiked above 0.7. Suggested cutoff could be used in future screenings.

### **Top Features Boxplots**



**Interpretation:** These clearly show deviation in key metrics between healthy and impaired speech, especially in pitch and word recall features.

## **8. API Design**

A reusable Python function was developed that:

* Accepts an audio file path
* Runs full preprocessing and feature extraction
* Returns a cognitive risk score using the Isolation Forest model

This is designed to be deployable via a REST API (Flask/FastAPI).

## **9. Key Achievements**

* Full automation of voice preprocessing pipeline
* Clinically inspired feature set
* Successfully flagged at-risk samples using unsupervised learning
* Visual insights aligned with clinical expectations
* Setup for real-time application with API-ready backend

## **10. Reflections**

I chose unsupervised models to preserve generality and interpretability. This project showed that even basic ML techniques, when paired with meaningful features, can uncover real-world cognitive risk patterns. The next phase would involve clinician validation and expansion to multi-lingual, multi-demographic datasets.