International Journal
Information Technology Applications (ITA)

Volume 14, Number 1, June 2025

# DETECTION OF NEURODEGENERATIVE DISEASES FROM SPEECH USING MULTIPLE CLASSES

**Barbora Jurkovcová, Eugen Ružický, Štefan Kozák,
Ján Lacko, Alfréd Zimmermann**

***Abstract:***

*Early detection of neurodegenerative diseases such as Alzheimer's disease (AD), mild cognitive impairment (MCI), and Parkinson's disease (PD) is crucial for timely intervention and improved patient outcomes. This study presents a machine learning framework for non-invasive diagnosis using voice signals recorded via smartphones. A machine learning model was developed, trained on the Slovak EWA-DB database, and evaluated on a large cohort consisting of patients with AD-MCI, PD, and healthy controls (HC). This model uses multi-class classification, which is more challenging than binary classification. The model achieved results with 93.5% accuracy, 93.6% sensitivity, and an F1 score of 88.5%, which are comparable to the results obtained from the EWA-DB for binary classification tasks. These results were achieved through thorough data preprocessing, including stratified sampling by age and diagnosis, class balancing through synthetic oversampling of minority classes, and dimensionality reduction through principal component analysis while preserving key information. We plan to apply the results independently of the multi-category classification of AD, MCI, and HC for non-invasive screening strategies in clinical practice. The proposed approach highlights the potential of speech biomarkers in combination with machine learning to improve early diagnosis across multiple classes.*

***Keywords:***

*Artificial intelligence, machine learning, smartphone application, EWA-DB database, detection of neurodegenerative diseases.*

## Introduction

Digital biomarkers represent a revolutionary tool in clinical decision-making, as they enable accurate, objective and continuous measurement of patients' health status using smartphones and wearable devices. These technologies can detect subtle motor, cognitive and speech changes that are often imperceptible by traditional methods. Speech signals such as changes in rhythm, pitch or articulation are showing promise in the diagnosis of neurological disorders such as Alzheimer's disease (AD) and Parkinson's disease (PD). Thanks to artificial intelligence and machine learning, these features can be analysed automatically, increasing accuracy and reducing the burden on clinical staff.

Alzheimer's disease is the most common type of dementia, and other health problems can manifest in addition to cognitive impairment. The process of predicting the disease is time-consuming and in the early stages of Alzheimer's disease, mild cognitive impairment (MCI) can be diagnosed as a very mild manifestation of the disease. Speech impairment is one of the earliest symptoms of AD patients. Many studies have demonstrated the potential of automated acoustic assessment using acoustic and linguistic features extracted from speech.

# 1 Analysis of Alzheimer's and Parkinson's Disease Prediction Based on Speech

In this section, we focus on significant advances and applications of machine learning in the detection of AD, MCI, and PD between 2000 and 2025, highlighting the best results achieved in the detection and prediction of these neurodegenerative diseases.

## 1.1 Analysis of Research for the Prediction of Alzheimer's Disease

Studies using automatic speech recognition (ASR) have focused on improving AD classification. The ADReSS (2020) and ADReSSo (2021) challenges evaluated speech-based cognitive decline detection [1, 2, 3]. ADReSS achieved 88% accuracy using only transcriptions without audio, and by combining audio with text features, accuracy increased to 93.8% [4]. Although manual transcriptions achieve the highest accuracy, ASR outputs were comparable to them [5]. A thoroughly tuned BERT model on ASR transcriptions achieved similar performance to manual transcriptions [6]. The study [7] compared three ASR systems (Google Speech, Whisper, and Wave2vec2), and Google achieved the highest accuracy of AUC 0.87.

The latest research on Alzheimer's disease and mild cognitive impairment [8] has focused more on the use of digital biomarkers for early diagnosis, reviewing 431 studies from five online databases. The research also focused on the use of automated speech analysis as a non-invasive biomarker for early diagnosis of AD in other languages. We focused mainly on studies with a larger number of AD cases. The study [9] analysed the acoustic characteristics of speech in 8779 Japanese seniors who read short sentences, and machine learning achieved a classification accuracy of AUC 0.61–0.77 in distinguishing MCI from healthy controls, with individuals with global cognitive impairment defined using the MMSE in the range of 20–23. In the publication [10], they applied a simple phonemic verbal fluency task in Thai to 100 individuals and proposed new feature extraction methods based on phonemic grouping and switching, achieving 65.5% accuracy in MCI classification.

Yamada et al. [11] compared speech responses to everyday questions in 94 Japanese participants with neuropsychological tasks and found that everyday questions can detect MCI with 87.4% accuracy. A study [12] of an NLP model based on simple picture description tasks in Slovak recorded using a smartphone achieved 95% accuracy in AD classification after balancing the data of healthy and AD individuals in a sample of 1117 people, of whom approximately 10% were diagnosed with AD and MCI. The study [13] used lexical-semantic and acoustic speech characteristics to identify MCI and AD in the early stages in participants from an English-speaking population in the US, with lexical-semantic scores achieving an AUC of 0.80 and correlating with amyloid-β biomarkers in cerebrospinal fluid. Ambrosini et al. [14] conducted a multilingual study in Italy and Spain with 133 participants, where automatic analysis of spontaneous speech enabled the classification of cognitive decline with 84% accuracy in participants with mild impairment. In their publication [15], they implemented a speech analysis algorithm in a Spanish-speaking population using five language tasks and machine learning, achieving an AUC of 0.93 and 88.4% accuracy in distinguishing MCI and dementia from healthy individuals.

## 1.2 Analysis of Research for the Prediction of Parkinson's Disease

Parkinson's disease, the second most common neurological disorder after AD, causes not only movement disorders but also speech disorders. Changes in speech, such as reduced volume, slurred speech, or changes in speed, can serve as early and non-invasive indicators of the disease.

Research in the field of PD diagnostics has shifted significantly towards the use of speech signals as a non-invasive biomarker. In a review study [16], the authors focused on the current state of research in the field of clinical decision-making using speech signal analysis in AD and PD. Based on selected clinical and technical criteria, an analysis of 72 selected studies on the latest developments in the field of early warning of PD based on speech was performed.

Zhang et al. [17] developed a mobile system for analysing linear and nonlinear dysphonia features, achieving 88.7% accuracy in a large Chinese cohort and deploying it via a mobile app, enabling practical telemedicine applications. In the US, a web-based PD screening framework with a short speech task using XGBoost model [18], reached an AUC of 0.753, with interpretable outputs via additive explanations. Vasquez-Correa et al. [19] integrated speech and movement signals into end-to-end deep learning, with models trained on smartphone and laboratory data achieving more than 92% accuracy. A Chinese study [20] using 3-second /a/ and /u/ vowels and a FastAI deep learning model reported 82% accuracy. Wang et al. [21] applied LASSO-based feature selection and ML classifiers, achieving 76% accuracy and highlighting articulation as more discriminative than phonation. A study [22] conducted as part of the EWA research project obtained data from a mobile application and balanced the originally unbalanced data set (117 PD vs. 944 HC), achieving the best PD detection result with an accuracy of 96.9%. An Italian study using SVM analysed voice across PD stages [23], identified clinical-instrumental correlations, quantified L-Dopa effects, and achieved 98.3% accuracy.

A Chinese study [24] using three PD datasets and advanced speech processing achieved up to 95% diagnostic accuracy, identifying key variable features. Hemmerling et al. [25] applied Vision Transformers to mel-spectrograms with explainable AI, reaching 89.8% accuracy and improving model interpretability. These findings confirm that combining high-quality voice data with advanced algorithms enables effective, accessible, and interpretable PD diagnostics.

## 2 Analysis of Data from the EWA-DB Database Using Machine Learning Methods

Phones and tablets are becoming increasingly important in predicting neurological diseases. As part of our collaboration on the **EWA project** [26] with AXON PRO and the Institute of Informatics of the Slovak Academy of Sciences in Bratislava, we used smartphones to record healthy and sick patients with AD, MCI, and PD. We created the EWA-DB database, from which we used data to analyse the detection of three classes: HC, AD-MCI, and PD.

### 2.1 EWA-DB Database for the Diagnosis of Alzheimer's and Parkinson's Disease

As part of the EWA project, the EWA-DB database was created to process speech parameter data and evaluate parameters to distinguish between the speech of healthy people and people with cognitive disorders. The database is supplemented and modified in line with the latest research, as is the resulting mobile application for smartphones [26]. EWA-DB is a Slovak database of voice recordings created for the purpose of researching the early diagnosis of neurodegenerative diseases, particularly Alzheimer's disease, mild cognitive impairment, and Parkinson's disease [27, 28]. The database contains various speech tasks: phonation of the vowel "a", diadochokinesis "pa-ta-ka", naming of objects and activities, and description of pictures.

The database also contains detailed demographic data, cognitive test results (e.g., MoCA), and information on meeting the inclusion criteria. Transcriptions were first automatically generated and then manually edited and annotated by trained annotators.

Annotations recorded various types of hesitations, unspoken sounds, unclear expressions, phonetic errors, and distracting sounds.

EWA-DB is publicly available through the ELDA platform and represents a significant contribution to the development of automated systems for the diagnosis of neurodegenerative diseases based on speech [27]. Due to its scope, quality, and linguistic specifics, it is suitable for research in the fields of artificial intelligence, linguistics, psychology, and clinical neurology.

### 2.2 Preprocessing and Feature Scaling

The extracted acoustic and linguistic features from EWA-DB are high-dimensional and exhibit non-uniform, heavy-tailed distributions. To ensure numerical stability of the models and reduce the influence of outliers, we applied **Quantile Transformer** scaling [29]. This transformation maps each feature to a target uniform or normal distribution, which reduces skewness and extreme values, thereby improving the convergence of linear classifiers.

The choice of Quantile Transformer was motivated by its demonstrated robustness in prior research [12, 22], where it consistently outperformed linear scaling approaches on imbalanced, speech-derived datasets. We also compare the quantile transformer with Min-Max and Max-Abs scalers. In our comparative experiments, alternative linear scalers such as Min-Max or Max-Abs resulted in lower performance, particularly in sensitivity and MCC. Therefore, Quantile Transformer was selected as the default scaling method throughout all models.

Principal component analysis (PCA) is a linear dimensionality reduction method that transforms the original variables into a new set of orthogonal components, ranked according to the amount of explained variability [30]. The algorithm involves standardising the data, calculating the covariance matrix, and finally projecting the data into a lower-dimensional space. PCA allows you to simplify the data structure, remove redundancy, and highlight the most important patterns in the data, which is especially useful in data preprocessing and classification.

### 2.3 Synthetic Minority Oversampling Technique

The dataset is strongly imbalanced, with healthy individuals substantially outnumbering patients with AD-MCI or PD. This imbalance poses a risk of bias in the classifier toward the majority class, which is why we used a similar approach to that used in studies [12, 22]. To address this issue, we employed Synthetic Minority Oversampling Technique (**SMOTE**) during the training phase [31]. SMOTE synthetically generates new samples of the minority classes (AD-MCI and PD) in feature space, which results in a more balanced class distribution.

By enriching the dataset with additional minority samples, this approach mitigates the risk of majority-class dominance and improves the model's ability to detect early neurodegenerative changes in at-risk participants.

### 2.4 Application of Machine Learning Algorithms to Speech Analysis from EWA-DB

Machine learning (ML) methods and algorithms are effective tools in the diagnosis and treatment of various neurodegenerative diseases. In the case of some of these diseases, it is possible to estimate the current condition of the patient by analysing speech signals using ML techniques. For this reason, it is essential to pay attention to data preparation and preprocessing, including the use of scaling and normalization techniques.

One of the primary objectives of this study is to design and implement preprocessing procedures tailored for speech-based diagnostic modelling. To this end, we employed the Stochastic Average Gradient Augmented algorithm (SAGA) that integrates the strengths of various gradient-based techniques [32].

SAGA is particularly well-suited for weakly convex problems, requiring no manual adjustments, and it automatically adapts to the convexity level of the objective function. Its fast linear convergence and scalability make it an effective choice for large and sparse datasets commonly encountered in machine learning applications.

### 2.5 Methodological Framework and Evaluation Metrics for Diagnostics

However, the success of these methods essentially depends on the quality of the input data processed by the selected algorithm. To solve complex diagnostic tasks in selected neurodegenerative diseases, such as Alzheimer's and Parkinson's disease, the research team EWA-RT has developed and verified a general methodological framework. It uses a system of support programs (in Python) with a modular, interactive, and adaptive structure that allows the assessment of the disease status based on input data (Fig.1). For a comprehensive and clinically meaningful evaluation of the diagnostic capabilities of the multidimensional class model, we used a set of performance metrics: accuracy, sensitivity, specificity, and F1 score, each of which reflects a different dimension of predictive reliability [33, 34]. The Matthews correlation coefficient (MCC) extends the binary form, allowing for the evaluation of multi-class classification. It remains robust even when classes are significantly imbalanced and is particularly suitable for tasks in biomedicine and bioinformatics, where multi-class predictions with uncertain data are common [35].

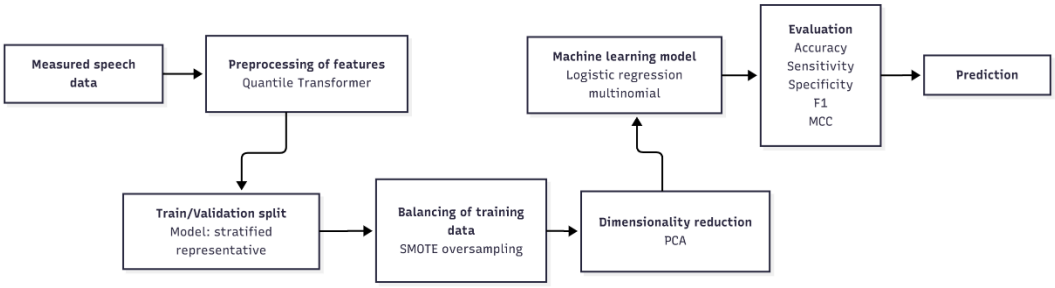



Fig.1. Block diagram of the ML modular system for detecting neurodegenerative diseases.

(Fig.1) presents a block diagram of the proposed machine learning pipeline for speech-based detection of Alzheimer's and Parkinson's diseases. The process begins with feature preprocessing and normalization using the Quantile Transformer to reduce noise and scale features. The dataset is then divided into training and validation subsets using a stratified representative split to preserve the class distribution. To address class imbalance, the SMOTE oversampling algorithm is applied exclusively to the training subset, ensuring that no synthetic samples are introduced into the validation data and thus preventing data leakage. Subsequently, dimensionality reduction is performed using Principal Component Analysis (PCA), followed by training a multinomial logistic regression model with SAGA optimization and L1 regularization. Model performance is evaluated using clinically relevant metrics, including Accuracy, Sensitivity, Specificity, F1-score, and the Matthews correlation coefficient (MCC). The final output represents a predictive model capable of classifying new speech recordings.

# 3 Classifying Neurodegenerative Diseases Based on Speech in Multiple Classes

To evaluate the effectiveness of machine learning for the early detection of neurodegenerative diseases from speech data, we developed and tested logistic regression–based classification model using the EWA-DB dataset. The dataset consists of high-dimensional acoustic and linguistic descriptors extracted from speech recordings, with participants labelled as healthy (HC), AD with mild cognitive impairment (AD-MCI), and Parkinson's disease (PD). Given the extreme dimensionality of the feature space (over 169 thousand features) and the inherent class imbalance (HC ≫ AD-MCI, PD), a standardized preprocessing pipeline was applied across both models to ensure comparability. Input features were scaled using a Quantile Transformer to normalize skewed distributions, and SMOTE oversampling was used to balance the minority classes (AD-MCI, PD) with the majority class HC. Dimensionality was then reduced with PCA, with thresholds between 82% and 95% of explained variance systematically evaluated. The model was trained using multinomial logistic regression with L1 regularization, optimized using the SAGA solver, which is suitable for high-dimensional problems with low density. The results shown in (Tab.3) showed the difference in training and testing accuracy as an additional indicator of stability to assess potential overfitting.

## 3.1 Stratified Selection of the Most Representative Samples

Our first idea was to build the validation set from the most representative cases; while keeping the more diverse/borderline cases for training so the model would learn robust decision boundaries. We therefore used a stratified split over two factors: diagnostic category (HC, AD-MCI, PD) and age bin (24–50, 51–60, 61–70, 71–80, 81–97). Each unique combination of diagnosis with age bin forms one stratum (e.g., “PD and 61–70”), (Tab.1), (Tab.2).

Within every stratum we measured the distance between samples in the scaled feature space and computed, for each sample, its average distance to all other samples in the same stratum. Samples with the smallest average distance (i.e., closest to the stratum centroid, thus most “typical”) were selected into the validation set; the remainder stayed in the training set. We selected 20% per stratum this way. The result is a validation set that faithfully reflects the population structure, and a training set enriched with more heterogeneous and boundary-case samples, which improves learning of discriminative class boundaries.

Table 1. Distribution of samples in the training and testing sets by diagnosis.

| Dataset | HC | MCI-AD | PD | Total samples |
|---|---|---|---|---|
| Training set | 823 | 96 | 129 | 1048 |
| Validation set | 204 | 24 | 31 | 259 |

Table 2. Distribution of samples in the training and testing sets by age bins.

| Dataset | 24-50 | 51-60 | 61-70 | 71-80 | 81-97 | Total samples |
|---|---|---|---|---|---|---|
| Training set | 25 | 290 | 398 | 243 | 92 | 1048 |
| Validation set | 7 | 73 | 98 | 59 | 22 | 259 |

The dataset was divided into a training set (1048 samples) and a validation set (259 samples). The distribution by diagnostic categories is summarized in (Tab.1), while the distribution by age bins is shown separately in (Tab.2). As shown in (Tab.1) and (Tab.2), the dataset was imbalanced, with healthy individuals forming the majority class. Most participants fell into the 51–70 age bins, while extreme bins (24–50 and 81–97 years) were sparsely represented. Preserving this demographic structure was essential, as speech features are strongly age-dependent and could bias classification if unevenly represented.

### 3.2 Machine Learning and Validation Workflow for Detecting AD-MCI and PD

To handle class imbalance, we applied SMOTE algorithm, which synthetically expanded the minority classes and yielded a balanced training set. Finally, we applied Principal Component Analysis (PCA) to reduce dimensionality. Out of the original more 169 thousand extracted features, PCA retained only 330 principal components while preserving 82% of the variance in the data (see (Tab.3). This dramatically reduced computational complexity and eliminated redundant or noisy features, while maintaining most of the clinically relevant information.

To further control overfitting in the high-dimensional feature space, the logistic regression model was trained with L1 regularization, which promotes sparsity by shrinking non-informative coefficients toward zero and thus performs embedded feature selection. For optimization we employed the SAGA algorithm, a stochastic variant of gradient descent that is particularly well-suited for large-scale and sparse problems. This combination enabled efficient training while focusing on the most discriminative speech features, further improving model generalization.

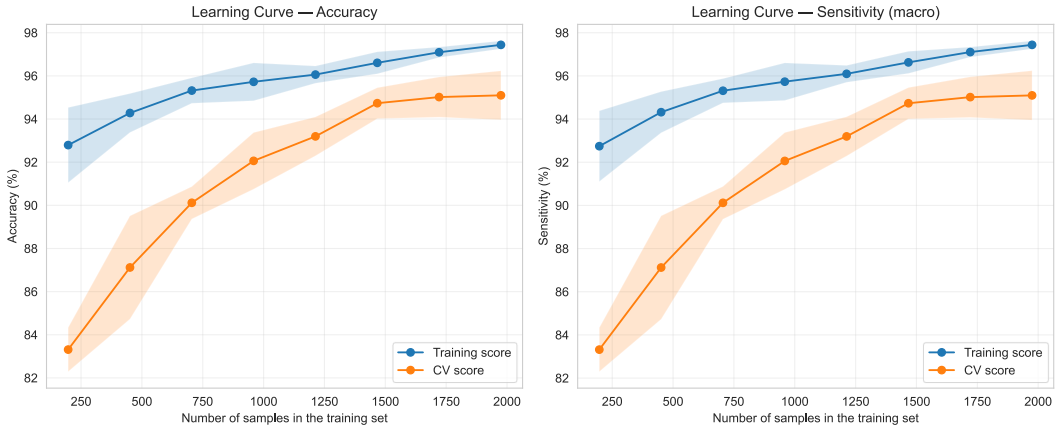



Fig.2. Cross-validation learning curves (accuracy and sensitivity).

To evaluate whether this training strategy leads to robust learning, we first inspected cross-validation learning curves (Fig.2). Both accuracy and macro-sensitivity increased steadily with larger training sets, while the gap between training and validation narrowed. This indicates that the model benefited from additional data and avoided severe overfitting, achieving good generalization capacity.

A complementary hold-out learning curve with the independent test set (Fig.3) further confirmed that test accuracy remained stable (approx. 92–93%) as the training size increased, supporting the stability and reproducibility of the model.

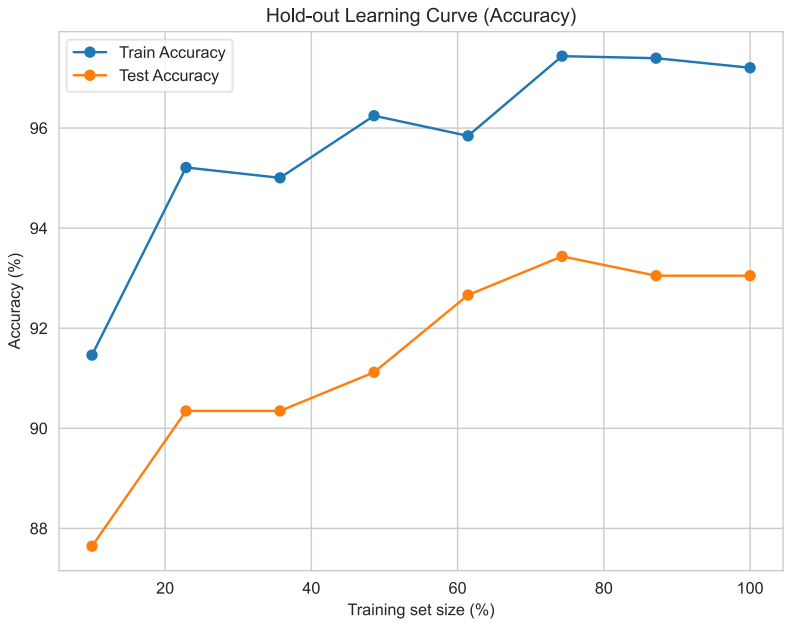



Fig.3. Hold-out learning curve on the independent test set.

The multinomial logistic regression model trained with this configuration achieved:

- Accuracy: 93.5%, Sensitivity: 93.6%, F1 Score: 88.5%
- MCC: 83.7%

(Fig.4) presents the confusion matrix for Model 1. The results demonstrate that the classifier performed consistently well across all three diagnostic categories.

- HC (class 0): The vast majority of samples (961) were correctly classified, with only a small proportion being misclassified.
- AD-MCI (class 1): Nearly all cases were correctly identified (115), with only 5 false negatives.
- PD (class 2): Most patients were classified correctly (146), with just a handful of errors.

Overall, the confusion matrix indicates a balanced and reliable performance across HC, AD-MCI, and PD participants. The strong diagonal dominance shows that the model not only detected the majority class accurately but also maintained high precision and recall for the minority classes, which are most critical in early neurodegenerative disease detection.

Receiver operating characteristic (ROC) analysis in (Fig.5) demonstrated high separability among all three classes. The area under the curve (AUC) was 0.989 for healthy controls, 0.996 for AD-MCI, and 0.988 for PD, with a macro-average AUC of 0.991. These findings further confirm that the classifier achieves both robust and clinically meaningful performance for the early detection of neurodegenerative disorders. The results indicate that representative stratified sampling, in combination with balanced training data and dimensionality reduction, enables reliable classification of neurodegenerative risk based on speech-derived features.

**Confusion Matrix**

| | | Predicted | | |
|---|---|---|---|---|
| | | **HC** | **AD** | **PD** |
| **Actual** | **HC** | 961 | 33 | 33 |
| | **AD** | 5 | 115 | 0 |
| | **PD** | 6 | 8 | 146 |

Fig.4. The confusion matrix shows results for three classes: HC, AD-MCI, PD.

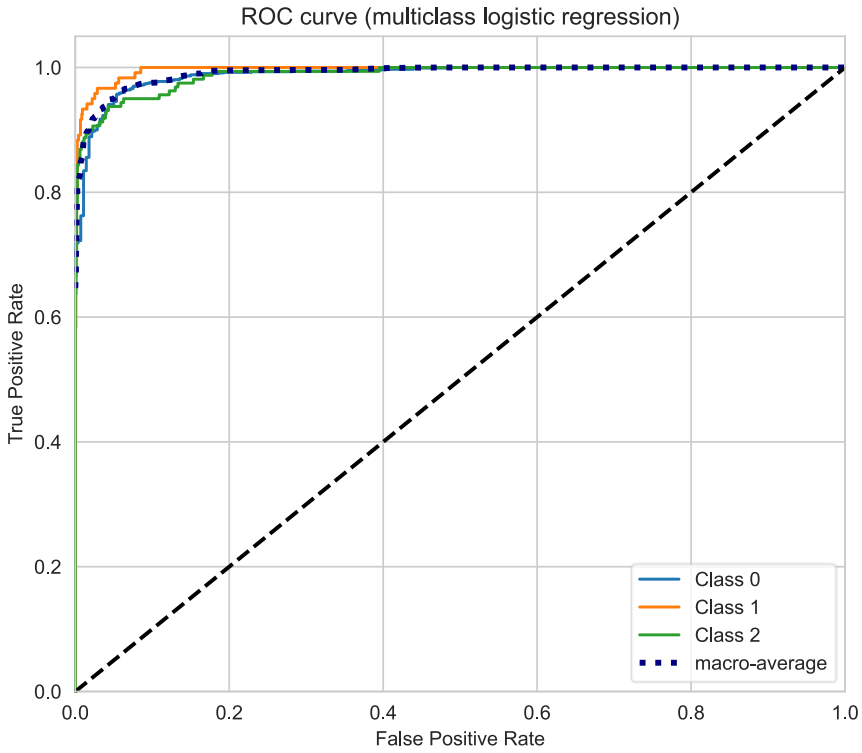



Fig.5. ROC curves for classes: Class 0 = HC, Class 1 = AD-MCI, Class 2 = PD.

### 3.3 Controlling Overfitting via PCA Variance Thresholds

To investigate the influence of dimensionality reduction on classification performance, we systematically varied the proportion of variance retained by PCA. Several thresholds were tested (70%, 75%, 82%, 85%, 88%, 90%, 92%, 95%), while keeping all other parameters constant (SMOTE oversampling, Quantile Transformer scaling, and logistic regression with L1 regularization).

For enhanced interpretability, (Tab.3) presents both the percentage of retained variance and the corresponding number of features (principal components) obtained after dimensionality reduction. This representation illustrates how the originally high-dimensional feature space was effectively compressed to a minimal set of features, while preserving most of the informative content in PCA %.

The primary evaluation criterion was performance on the independent test set, assessed by Accuracy, Sensitivity, F1 score, and MCC. We also considered the difference between training and testing accuracy, as this reflects the degree of overfitting or underfitting. The optimal configuration should therefore combine strong testing performance with a limited difference between training and testing accuracy.

As summarized in (Tab.3), the most advantageous Accuracy difference between training and testing (4.5 and 4.8, respectively) was observed at 82% and 85% PCA, as it optimizes the lowest risk of overfitting while maintaining competitive performance across all metrics. Although higher PCA thresholds (88%–92%) yielded slightly better F1 and MCC scores in the complete dataset, they also showed a larger difference between training and testing of more than 5%, indicating overfitting. Based on the balance between robust testing performance and model stability, PCA variance retention at 82% was chosen as the final model configuration.

Table 3. PCA is reported as retained variance (%)
and the resulting number of features and corresponding metrics in %.

| PCA % | No. features | Accuracy Train-Test Gap | Sensitivity | F1 Score | MCC |
|---|---|---|---|---|---|
| 70 % | 175 | 3.5 | 90.3% | 83.2 | 75.7 |
| 75 % | 223 | 4.2 | 92.3 | 85.8 | 79.6 |
| 82 % | 330 | 4.5 | 93.6 | 88.5 | 83.7 |
| 85 % | 396 | 4.8 | 94.1 | 89.4 | 85 |
| 88 % | 478 | 5.1 | 94.4 | 90.4 | 86.3 |
| 90 % | 544 | 5.6 | 94.2 | 90.3 | 86.2 |
| 92 % | 619 | 5.6 | 94.3 | 90.6 | 86.6 |
| 95 % | 715 | 5.7 | 94.3 | 90.6 | 86.6 |

### 3.4 Alternative Sampling and Scaling Methods

In addition to evaluating different PCA thresholds, we systematically compared alternative data preprocessing strategies for the proposed model. The default pipeline combined SMOTE oversampling with Quantile Transformer scaling. To assess the robustness of this configuration, we substituted SMOTE with ADASYN for synthetic minority oversampling and replaced the Quantile Transformer with linear scaling methods (Min-Max and Max-Abs). All other pipeline components, including the logistic regression classifier and PCA-based dimensionality reduction, remained unchanged.

As summarized in (Tab.4), none of the alternative pipelines outperformed the default configuration. Specifically, ADASYN consistently resulted in lower sensitivity and MCC values, indicating reduced effectiveness in capturing minority classes. Similarly, Min-Max and Max-Abs scaling underperformed compared to the Quantile Transformer, underscoring the importance of nonlinear feature transformation for the highly skewed distributions present in the EWA-DB speech

data. These findings confirm that the combination of SMOTE oversampling and Quantile Transformer scaling provides the most reliable preprocessing strategy for simultaneous AD-MCI and PD detection.

Table 4. Comparison of alternative sampling and scaling methods for the model.

| Sampling | Scaling | Accuracy | Sensitivity | F1 Score | MCC |
|---|---|---|---|---|---|
| ADASYN | Min-Max | 84.3% | 85.4% | 76.4% | 65.3% |
| ADASYN | Max-Abs | 83.9% | 87.4% | 76.4% | 66.0% |
| SMOTE | Min-Max | 83.1% | 84.5% | 74.9% | 63.4% |
| SMOTE | Max-Abs | 84.9% | 85.4% | 76.7% | 66.3% |

## 4 Discussion

In recent years, several solutions have been developed in the field of neurodegenerative disease detection, with the EWA-RT research team achieving comparable or higher accuracy than published studies. The discussion includes a comparison of selected results from the EWA-DB database with existing solutions, with the EWA-RT team's approach representing a new method of multi-class AD and PD detection based on linguistic and lexical characteristics from the EWA-DB database. The variability of speech tasks, extensive recordings, and detailed annotations provide a solid basis for evaluating algorithms and methodologies, enabling results comparable to international research.

The similar study from 2024 presented an automated speech analysis algorithm for detecting cognitive impairment (MCI and dementia) in Spanish-speaking individuals [15], Participants completed four short speech tasks, picture description ("Cookie theft"), phonemic fluency (words beginning with "F"), alternating semantic fluency (fruit and sports), and semantic fluency (animals), via a web or mobile application in less than five minutes. The recordings were processed using speech-to-text transcription and digital signal processing to extract acoustic and linguistic features. The final machine learning model achieved high classification performance: accuracy of 87.5–88.4%, sensitivity of 83.3–87.5%, and specificity of 89.2%.

Using the proposed methodology, the EWA-RT team achieved excellent classification results from the EWA-DB dataset, distinguishing patients with Alzheimer's disease, mild cognitive impairment (AD-MCI), and Parkinson's disease (PD) from healthy control subjects. The model achieved an accuracy of 93.3%, with a specificity of 93.5%, an F1 score of 88.4%, and a Matthews correlation coefficient of 83.2%. These results were enabled by rigorous data preprocessing, including stratified sampling by age and diagnosis, noise reduction and overfitting mitigation techniques, class balancing using SMOTE, and dimensionality reduction via principal component analysis while preserving key information.

Concurrently, multiple studies have examined the utility of the EWA-DB database for the prediction of Parkinson’s disease (PD). In 2022, the EWA-RT team demonstrated the effectiveness of machine learning in speech-based neurocognitive assessment of PD, achieving a classification accuracy of 96.9%, specificity of 89.3%, and Matthews correlation coefficient (MCC) of 83.2% [22]. The best results were achieved using linear regression in combination with Quantile Transformer scaling.

Further research explored cross-linguistic differences between Spanish and Slovak in the selection of acoustic features relevant to PD detection, utilizing data from the EWA-DB and PC-GITA datasets [36, 37]. The highest predictive performance within EWA-DB was obtained by combining optimized feature sets with spontaneous speech, resulting in an accuracy of 69.6% [36].

Additionally, the BDHPD study [37] introduced a novel deep learning architecture designed to enhance language generalization through self-supervised representation learning and transformation mechanisms. In its best single-language configuration, BDHPD achieved an accuracy of 83.6% and an F1 score of 70.3%.

Future work should focus on several directions to improve the EWA-DB framework. Extending the approach to preliminary classification by including Alzheimer's disease and mild cognitive impairment as separate classes alongside healthy controls would im-prove differentiation at an early stage. Similarly, distinguishing early-stage Parkinson's disease (PD) from HC is essential for early intervention. Combining linguistic characteristics with biomarkers such as neuroimaging, acoustic data, and cognitive scores could in-crease robustness and interpretability. The incorporation of longitudinal modelling to capture the temporal dynamics of disease progression (HC → MCI → AD or PD) would support predictive diagnostics and personalized treatment strategies.

## Conclusion

Future research should focus on expanding the EWA-DB database with long-term records of healthy individuals and patients, which will significantly advance the prediction of non-invasive diagnostic methods. The automatic speech database creation system allows speech to be collected from individuals in their home environment using common devices such as smartphones or laptops. This approach supports the automatic processing of voice responses in any language, including Slovak, and significantly reduces the cost and time required to create specialized speech databases. Such processing will enable rapid and longitudinal analysis of clinical data, allowing for disease prediction, personalized approach, and therapy design, thereby minimizing adverse effects. Ultimately, clinical interpretability remains essential for real-world application. The development of explainable AI techniques that provide transparent and clinically meaningful insights will be key to building trust and integration into healthcare workflows.

## Acknowledgement

This paper is supported by project GAAA/2023/6 “Basic research of new innovative methods using virtual reality for early detection of neurodegenerative diseases” and GAAA/2024/27 “Research of algorithms for process modelling, control and visualisation in domains of applied informatics”, Grant Agency Academia Aurea (GAAA) https://www.gaaa.eu/.

## Abbreviations

| | |
|---|---|
| AD | Alzheimer’s Disease |
| ADASYN | Adaptive Synthetic Sampling |
| ADReSS | Alzheimer’s Dementia Recognition through Spontaneous Speech |
| ADReSSo | Alzheimer’s Dementia Recognition through Spontaneous Speech only |
| AI | Artificial Intelligence |
| ASR | Automatic Speech Recognition |
| AUC | Area Under the Curve |
| BDHPD | Bilingual Dual-Head Parkinson’s Disease Model |
| EWA | project: Early Warning of Alzheimer |
| EWA-DB | Early Warning of Alzheimer Database |

| | |
|---|---|
| EWAT | Faculty of Informatics: EWA - Team |
| F1 | F1 Score (Harmonic Mean of Precision and Recall) |
| HC | Healthy Controls |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| MCC | Matthews Correlation Coefficient |
| MCI | Mild Cognitive Impairment |
| ML | Machine Learning |
| MMSE | Mini-Mental State Examination |
| MoCA | Montreal Cognitive Assessment |
| PCA | Principal Component Analysis |
| PD | Parkinson's Disease |
| SAGA | Stochastic Average Gradient Augmented |
| SMOTE | Synthetic Minority Oversampling Technique |
| SVM | Support Vector Machine |
| Wave2vec2 | Wav2Vec 2.0 (Self-Supervised Speech Representation Model) |
| XAI | Explainable Artificial Intelligence |
| XGBoost | Extreme Gradient Boosting |

## References


[1] Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., MacWhinney, B.: Alzheimer's dementia recognition through spontaneous speech. Front. Comput. Sci. 2021, 3, 780169.

[2] Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B.: Detecting cognitive decline using speech only: The ADReSSo challenge. arXiv 2021, arXiv:2104.09356.

[3] Ding, K., Chetty, M., Noori Hoshyar, A., Bhattacharya, T., Klein, B.: Speech-based detection of Alzheimer's disease: A survey of AI techniques, datasets and challenges. Artif. Intell. Rev. 2024, 57(12), 325.

[4] Martinc, M., Haider, F., Pollak, S., Luz, S.: Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. Front. Aging Neurosci. 2021, 13, 642647.

[5] Soroski, T., da Cunha Vasco, T., Newton-Mason, S., Granby, S., Lewis, C., Harisinghani, A., et al.: Evaluating web-based automatic transcription for Alzheimer speech data: Transcript comparison and machine learning analysis. JMIR Aging 2022, 5(3), e33460.

[6] Li, C., Cohen, T., Pakhomov, S.: The far side of failure: Investigating the impact of speech recognition errors on subsequent dementia classification. arXiv 2022, arXiv:2211.07430.

[7] Heitz, J., Schneider, G., Langer, N., Calzolari, N., Kan, M.Y., Hoste, V., et al.: The influence of automatic speech recognition on linguistic features and automatic Alzheimer's disease detection from spontaneous speech. In: Proceedings of the International Conference on Computational Linguistics, ELRA and ICCL: 2024, pp. 15955–15969.

[8] Qi, W., Zhu, X., Wang, B., et al.: Alzheimer's disease digital biomarkers multidimensional landscape and AI model scoping review. npj Digit. Med. 2025, 8, 366. https://doi.org/10.1038/s41746-025-01640-z.

[9] Nagumo, R., Zhang, Y., Ogawa, Y., Hosokawa, M., Abe, K., Ukeda, T., et al.: Automatic detection of cognitive impairments through acoustic analysis of speech. Curr. Alzheimer Res. 2020, 17(1), 60–68.

[10] Metarugcheep, S., Punyabukkana, P., Wanvarie, D., Hemrungrojn, S., Chunharas, C., Pratanwanich, P.N.: Selecting the most important features for predicting mild cognitive impairment from Thai verbal fluency assessments. Sensors 2022, 22(15), 5813. https://doi.org/10.3390/s22155813.

[11] Yamada, Y., Shinkawa, K., Nemoto, M., Ota, M., Nemoto, K., Arai, T.: Speech and language characteristics differentiate Alzheimer's disease and dementia with Lewy bodies. Alzheimer's Dement. Diagn. Assess. Dis. Monit. 2022, 14(1), e12364.

[12] Képešiová, Z., Ružický, E., Kozák, Š., Malaschitz, R., Zimmermann, A.: Application of advanced machine learning algorithms for early detection of mild cognitive impairment and Alzheimer's disease. In: Proceedings of the 2023 International Scientific Conference on Computer Science (COMSCI), IEEE: Sozopol, Bulgaria, 2023, pp. 1–5. https://doi.org/10.1109/COMSCI59259.2023.10315946.

[13] Hajjar, I., Okafor, M., Choi, J.D., Moore, E., Abrol, A., Calhoun, V.D., Goldstein, F.C.: Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. Alzheimer's Dement. Diagn. Assess. Dis. Monit. 2023, 15(1), e12393.

[14] Ambrosini, E., Giangregorio, C., Lomurno, E., Moccia, S., Milis, M., Loizou, C., Ferrante, S.: Automatic spontaneous speech analysis for the detection of cognitive functional decline in older adults: Multilanguage cross-sectional study. JMIR Aging 2024, 7, e50537. https://doi.org/10.2196/50537.

[15] Kaser, A.N., Lacritz, L.H., Winiarski, H.R., Gabirondo, P., Schaffert, J., Coca, A.J., Cullum, C.M.: A novel speech analysis algorithm to detect cognitive impairment in a Spanish population. Front. Neurol. 2024, 15, 1342907. https://doi.org/10.3389/fneur.2024.1342907.

[16] De Silva, U., Madanian, S., Olsen, S., Templeton, J., Poellabauer, C., Schneider, S., Narayanan, A., Rubaiat, R.: Clinical decision support using speech signal analysis: Systematic scoping review of neurological disorders. J. Med. Internet Res. 2025, 27, e63004. https://doi.org/10.2196/63004.

[17] Zhang, L., Qu, Y., Jin, B., Jing, L., Gao, Z., Liang, Z.: An Intelligent Mobile-Enabled System for Diagnosing Parkinson Disease: Development and Validation of a Speech Impairment Detection System. JMIR Med Inform. 2020, 8(9), e18689. https://doi.org/10.2196/18689

[18] Rahman, W., Lee, S., Islam, M.S., Antony, V.N., Ratnu, H., Ali, M.R., Hoque, E.: Detecting Parkinson Disease Using a Web-Based Speech Task: Observational Study. J Med Internet Res. 2021, 23(10), e26305. https://doi.org/10.2196/26305

[19] Vasquez-Correa, J.C., Arias-Vergara, T., Klumpp, P., Pérez-Toro, P.A., Orozco-Arroyave, J.R., Nöth, E.: End-to-End Modeling of Speech and Gait from Patients with Parkinson's Disease: Comparison between High Quality vs. Smartphone Data. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: Toronto, Canada, 2021, pp. 7298–7302.

[20] Tandjung, M.D., Wu, J.C.M., Wang, J.C., Li, Y.H.: An Implementation of FastAI Tabular Learner Model for Parkinson's Disease Identification. In: Proceedings of the 2021 9th International Conference on Orange Technology (ICOT), IEEE: Tainan, Taiwan, 2021. https://doi.org/10.1109/ICOT54518.2021.9680650

[21] Wang, Y., Li, X., Chen, H., Zhang, J.: Automatic Detection of Parkinson's Disease Using Chinese Speech Signals and Feature Selection Algorithms. J Biomed Eng. 2022, 39(2), 145–158.

[22] Képešiová, Z., Kozák, Š., Ružický, E., Zimmermann, A., Malaschitz, R.: Comparative Analysis of Advanced Machine Learning Algorithms for Early Detection of Parkinson Disease. Proceedings 2022, Cybernetics & Informatics (K&I), Visegrád, Hungary, pp. 1-6. https://doi.org/10.1109/KI55792.2022.9925942

[23] Suppa, A., Costantini, G., Asci, F., Di Leo, P., Al-Wardat, M.S., Di Lazzaro, G., Saggio, G.: Voice in Parkinson's Disease: A Machine Learning Study. Front. Neurol. 2022, 13, 831428. https://doi.org/10.3389/fneur.2022.831428.

[24] Yuan, L., Liu, Y., Feng, H.-M.: Parkinson disease prediction using machine learning-based features from speech signal. Service Oriented Computing and Applications, 18(1), 2024, 101-107. doi:10.1007/s11761-023-00372-w.

[25] Hemmerling, D., Zakrzewski, M., Wodzinski, M., Dudek, M., Gaciarz, F., Wojcik-Pedziwiatr, M., Rumezhak, T.: Improving AI Interpretability for Multilingual Parkinson's Disease Classification through Voice Analysis. In AAAI Bridge Program on AI for Medicine and Healthcare, PMLR: Palo Alto, CA, USA, 2025, pp. 49–55.

[26] **Project EWA. Home.**
Available online: https://www.projektewa.sk/en/index.html (accessed on 24 August 2025).

[27] Rusko, M. et al.: EWA-DB Early Warning of Alzheimer speech database. https://catalog.elra.info/en-us/repository/browse/ELRA-S0489/, (2023).98. EWA-DB – Early Warning of Alzheimer speech database. https://zenodo.org/records/10952480, https://doi.org/10.5281/zenodo.10952480, 2024.

[28] Rusko, M., Sabo, R., Trnka, M.; Zimmermann, A., Malaschitz, R., Ružický, E. et al.: Slovak database of speech affected by neurodegenerative diseases. Sci Data 11, 1320, 2024. https://doi.org/10.1038/s41597-024-04171-6.

[29] Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. Bioinformatics 2003, 19, 185–193. https://doi.org/10.1093/bioinformatics/19.2.185.

[30] Jolliffe, I.T., Cadima, J.: Principal Component Analysis: A Review and Recent Developments. Philos. Trans. R. Soc. A 2016, 374, 20150202. https://doi.org/10.1098/rsta.2015.0202

[31] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002, 16, 321–357. https://doi.org/10.1613/jair.953

[32] Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives. Advances in Neural Information Processing Systems, 2014.

[33] Sokolova, M., Lapalme, G.: A Systematic Analysis of Performance Measures for Classification Tasks. Inf. Process. Manag. 2009, 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

[34] Takahashi, K., Yamamoto, K., Kuchiba, A., Koyama, T.: Confidence Interval for Micro-Averaged F1 and Macro-Averaged F1 Scores. Applied Intelligence 2022, 52, 4961–4972. https://doi.org/10.1007/s10489-021-02635-5.

[35] Gorodkin, J.: Comparing Two K-Category Assignments by a K-Category Correlation Coefficient. Computational Biology and Chemistry 2004, 28, 367–374. https://doi.org/10.1016/j.compbiolchem.2004.09.006.

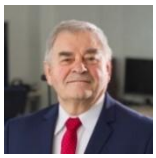
[36] Rekdal, M.: Speech Analysis for Automatic Parkinson's Disease Detection: Feature, Data, and Language Analysis for Performance Optimization. Master's Thesis, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2024.

[37] La Quatra, M., Orozco-Arroyave, J.R., Siniscalchi, M.S.: Bilingual Dual-Head Deep Model for Parkinson's Disease Detection from Speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2025, pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889445.

## Authors

**Mgr. Barbora Jurkovcová**
AXON PRO Ltd. Bratislava, Slovakia
jurkovcova@axonpro.sk
Her research interests include Applied informatics, modeling, visualization, and applications in medicine.

**Assoc. prof. RNDr. Eugen Ružický, PhD.**
Faculty of Informatics, Pan-European University, Bratislava, Slovakia
eugen.ruzicky@paneurouni.com
His research interests include Applied informatics, system analysis, modeling, visualisation and applications in medicine.



**prof. Ing. Štefan Kozák, PhD.**
Faculty of Informatics,
Pan-European University in Bratislava, Slovakia
stefan.kozak@paneurouni.com
His research interests include system theory, linear and nonlinear control methods, numerical methods, algorithms and software for modeling, control, signal processing, IoT, and embedded intelligent systems applications in a wide range of industrial, healthcare, banking and service sectors.

**RNDr. Ján Lacko, PhD.**
Faculty of Informatics, Pan-European University, Bratislava, Slovakia
jan.lacko@paneurouni.com
His research interests include digitization of objects from the field of cultural heritage, healthcare, industry, urban planning. and their display by various techniques, including virtual and augmented reality.

**RNDr. Alfréd Zimmermann**
AXON PRO Ltd. Bratislava, Slovakia
zimmermann@axonpro.sk
He is owner and CEO of the company. Professional and scientific interests include artificial intelligence in general and natural language processing.