Technical documentation Parkinson's Voice Based Risk Predictor

Anne Fernando
Ruxandra Speranta Oprea

Table of Contents

Device Description and Specification

General Description of the AI-Based Device

This software application is an AI diagnostic support tool developed to assist licensed healthcare professionals in the early detection of Parkinson's disease. It analyses structured voice data based on 22 clinically relevant acoustic features, which are extracted from patient's voice recordings. The tool generates a binary prediction (presence or absence of Parkinson's risk) using a pre-trained XGBoost classifier and SHAP-based interpretability. The interface is deployed using Streamlit.

Intended Purpose

- **Purpose:** To support the diagnostic process of Parkinson's disease by offering clinicians a probabilistic assessment of risk based on non-invasive vocal features.
- **Indications:** The tool is intended for patients presenting with speech or motor disturbances suggestive of early Parkinson's disease or under clinical follow-up for neurodegenerative evaluation.
- **Contraindications:** The device is not to be used as a stand-alone diagnostic instrument. It is intended to be used in conjunction with other clinical assessments.
- **Warnings:** The output of this tool is to be interpreted by qualified medical professionals. It should not be used without clinical judgment or outside the context of comprehensive medical evaluation.

- **Patient Population:** Adults at risk of Parkinson's disease, particularly those in the elderly demographic or with symptoms of neurological decline.
- **Medical Conditions Addressed:** Parkinson's disease and related motor and speech disorders.

AI Interaction

- **User Interface:** A Streamlit-based graphical interface allows clinicians to input data manually or upload patient's CSV files containing biomarker values.
- **ML Backend:** The app loads a pre-trained XGBoost model and StandardScaler for real-time prediction.
- **Interpretability:** SHAP (SHapley Additive exPlanations) visualizations provide transparency and support human-in-the-loop decision-making.
- **Data Handling:** Predictions and inputs are logged locally under a unique patient ID and timestamp.
- **External Systems:** The tool does not connect to electronic health records or other AI tools, maintaining a secure and isolated local execution environment.

Technical Specification

- Input Format: 22 standardized acoustic biomarkers extracted from voice recordings.
- Output: Binary risk classification (Parkinson's detected / No Parkinson's detected) and associated confidence score.
- Model Used: XGBoost binary classifier, trained on a validated public dataset.
- Technologies: Python 3.8+, Streamlit, scikit-learn, SHAP, pandas, NumPy.

- Execution Requirements: Low-compute hardware (laptop/server with basic CPU); compatible with offline execution.
- GUI Features: Welcome screen, login system with patient ID entry, data input page (CSV or form), result display with SHAP visualization, and logout functionality.

Designation / Classification

- Device Category: Software as a Medical Device (SaMD)
- MDR Classification: Class IIa under Rule 11 of MDR Annex VIII

This classification applies because the software is explicitly intended to provide information used by healthcare professionals to support diagnostic decisions.

Justification:

- The software processes medical data
- Its output contributes directly to diagnosis
- It is to be used by qualified clinicians
- AI Act Classification: High-risk AI system, as defined by the proposed EU AI Act (Annex III, paragraph 5), due to its role in medical decision support.

Disclaimer: While the device is built with the design principles of a regulated diagnostic support tool, it remains a student prototype not yet CE marked or validated in clinical trials.

Design and Manufacturing Information

Description of the Design

- The device was built in modular steps:
- Dataset selection and feature engineering using the UCI Parkinson's dataset.
- Training of a robust and interpretable ML model (XGBoost), alongside data standardization with StandardScaler.

- GUI development with Streamlit to facilitate user input and feedback.
- Integration of SHAP for model transparency.
- Creation of secure logging features linked to patient ID and timestamp.
- The modularity, allows for easy replacement or update of the model and GUI components.

Description of the AI Development

1. Development Methods and Integration of External Tools

The AI system was developed using open-source tools and libraries, with no pre-trained model used. Instead, a XGBoost model was trained specifically for Parkinson's detection based on voice features.

The development process included:

- Exploratory data analysis of the UCI Parkinson's dataset
- Preprocessing and standardization using Scikit-learn's StandardScaler
- Model selection, parameter tuning, and training using XGBoost

XGBoost was selected as the ML model because it offered a good balance between accuracy, interpretability, and ease of use. The dataset is a tabular collection of 195 voice recordings with 22 features, and the model performs well on structured datasets and includes built-in regularization to reduce the risk of overfitting, which is especially important in medical data analysis where samples are limited.

Another key reason for choosing XGBoost was its compatibility with SHAP, a library that helps explain model predictions. Since the goal of this app is to support healthcare professionals rather than replace them, having transparent and interpretable predictions is important. SHAP explanations

allowed us to show which features contributed most to a particular prediction, helping clinicians make more informed decisions.

Alternative models such as logistic regression and support vector machines were also considered. However, these either lacked the flexibility or the interpretability required for the project. Deep learning models were ruled out due to the small dataset size and their complexity, which would make them harder to validate and explain in a clinical setting.

- Integration of SHAP (SHapley Additive exPlanations) for model transparency
- Development of a user interface in Streamlit to operationalize the AI pipeline

2. Design Specifications and Algorithmic Logic

The system is designed as a binary classification model to predict the presence or absence of Parkinson's disease based on acoustic biomarkers.

- Algorithm: XGBoost (Extreme Gradient Boosting) was chosen for its high accuracy, robustness to overfitting, and feature importance interpretability.
- Classification Objective: The model is trained to optimize the binary log loss (classification with probability), maximizing discrimination between Parkinson's and healthy control samples.

Key design choices and rationale:

- Use of SHAP for interpretable predictions to fulfill transparency and oversight obligations
- Input limited to 22 validated acoustic biomarkers to ensure data consistency and medical relevance
- Exclusion of personal identifiers to ensure data minimization

Assumptions:

- The model is used by licensed healthcare professionals who understand the limitations of ML predictions
- The model is not intended to replace traditional diagnostic procedures, but to assist clinicians in early decision-making

3. System Architecture

The architecture is modular and designed for transparency and flexibility:

- Frontend: Streamlit interface for manual data entry or CSV upload
- Preprocessing Layer: Scikit-learn StandardScaler loaded from .pkl
- Inference Engine: XGBoost classifier loaded from .pkl
- Interpretability Module: SHAP explainer invoked only upon user request
- Logging Module: Stores predictions and inputs by patient ID with timestamp
- All modules run locally, requiring minimal hardware and no internet connection
- Computational resources: Developed and tested on a consumer-grade laptop with an Intel i5 CPU and 8GB RAM; no GPU required

4. Data Requirements and Handling

- Dataset: UCI Parkinson's dataset (publicly available, 195 instances, 22 acoustic features per sample)
- Provenance: Collected by Max Little et al. at Oxford University and available from the UCI Machine Learning Repository

Data Characteristics:

- 31 patients (23 with Parkinson's, 8 healthy controls)
- Each patient contributed multiple voice recordings

Preprocessing:

- No missing values
- Data scaled using StandardScaler
- No imputation needed
- No class rebalancing applied due to manageable imbalance

Labeling:

- Supervised binary labels: status = 1 (Parkinson's), status = 0 (Healthy)

5. Human Oversight and Interpretability Measures (Article 14)

The app is built to facilitate clinician involvement in every step:

- Predictions are accompanied by confidence scores
- SHAP plots show feature contributions for transparency
- Results are displayed in natural language for interpretability (e.g. "Parkinson's detected" with 87% confidence)
- SHAP interpretability allows human users to audit how the AI reached its conclusions, promoting trust and accountability
- No autonomous decision-making is permitted; all decisions remain with the human user

6. Predetermined Changes and Lifecycle Planning

Predefined Changes:

- Model and scaler can be updated manually by the developer through retraining
- Future updates may include additional training data or performance monitoring tools

Technical Safeguards:

- Separation of data preprocessing and model files supports modular updates
- Versioning of .pkl files ensures reproducibility

7. Validation and Testing Procedures

Testing Environment:

- All components were tested locally in an offline environment

Cross-validation:

- The model was evaluated using 10-fold cross-validation
- Average accuracy exceeded 85%

Metrics Evaluated:

- Accuracy, precision, recall, F1-score, ROC-AUC

Robustness Checks:

- Model behavior tested on edge-case inputs
- SHAP explanations checked for logical consistency

Discrimination and Bias:

- Dataset does not contain demographic information (e.g. gender, age), so fairness cannot be evaluated directly
- However, SHAP analysis was used to ensure model is not overly sensitive to any single feature

Logs:

- Prediction results, inputs, and patient ID are stored locally in .csv for future audit and performance tracking

8. Cybersecurity Measures

As the app operates locally, there is no transmission of data over the internet

- All patient-related logs are saved to disk under user control
- No external API access is used
- Future plans may include encrypted logging or authentication mechanisms for shared devices

General Safety and Performance Requirements

The device, aClass IIa medical device, meets applicable General Safety and Performance Requirements (GSPR) as outlined in Annex I of EU Regulation

2017/745 (MDR). While this tool is not CE-marked, the development and documentation simulate compliance as expected in a regulated environment.

1. Justification of Applicability and Inapplicability

- Performance Requirements: The device performs reliably under intended use scenarios, producing consistent outputs.
- Risk Reduction: Risk controls such as disclaimers, input validation, confidence scoring, and interpretability are implemented.
- User Interface Safety: Clear and simple user interface minimizes misuse.

Inapplicable requirements include:

- Sterility and Biocompatibility: Not applicable, as the software is non-invasive.
- Implantable Devices: Not applicable.
- Radiation or Electrical Emission: Not applicable.

2. Reference to Applied Standards and Common Specifications

The following standards and frameworks were used to guide development:

ISO 14971:2019 Medical Devices Risk Management, used to assess and minimize system-level risks
IEC 62304:2006 Software Lifecycle for Medical Devices - Simulated adherence to basic principles (version control, modular design)
GDPR (EU 2016/679) Data Protection - Local-only data storage, no personally identifiable data processed
AI Act Proposal (2021/0106) High-risk AI Systems- Human oversight and transparency measures via SHAP

3. Controlled Documents and Records

The following documentation is available to support traceability and system performance:

- Model artifacts: model/xgb_model.pkl and model/scaler.pkl with known creation dates
- Logs: Each patient interaction stored in logs/{patient_id}_{timestamp}.csv
- Development Notes: Stored in Jupyter notebooks and version-controlled Python scripts
- Testing Outputs: Internal cross-validation scores logged for performance review

4. Evaluation of Requirements Fulfilment

Each applicable GSPR was evaluated against the implementation status of the prototype:

GSPR Requirement

Device must achieve intended performance, this was done by scoring an accuracy >85%, deterministic logic, SHAP explanations

Risks must be reduced as far as possible, done with the use of input validation, user prompts, logout safety, confidence display

Device must not compromise health or safety, achieved because the device is used only under professional supervision, and has a non-invasive design

Device must be usable as intended, the device has a Streamlit GUI, for which minimal training is required

Manufacturer must demonstrate performance, the model and scaler performance is logged and reproducible

Instructions must be clear, the app includes clear page-by-page instructions and interface labels

AI Performance Requirements

1. Functioning and Performance Capabilities

- The AI system is based on a supervised XGBoost classifier, trained on a validated dataset of structured voice features (UCI Parkinson's dataset).
- The system performs binary classification (Parkinson's / No Parkinson's) and outputs a confidence probability for each prediction.
- During internal validation, the model consistently achieved over 85% accuracy through 10-fold cross-validation.
- SHAP is used to provide visual explanations of feature contributions, improving transparency and understanding.

2. Intended Use Population and Limitations

- The system is designed for use by licensed healthcare professionals assessing adult patients with potential early symptoms of Parkinson's disease.
- The original dataset does not contain demographic markers (age, gender, ethnicity), so performance differences across population groups cannot currently be evaluated.

Limitations:

- Small dataset size (n=195) may restrict generalizability
- The model is sensitive to incomplete or misformatted input
- Performance may degrade on voice recordings obtained from different hardware or environments unless retrained

3. Foreseeable Unintended Outcomes and Risks

- False positives may lead to unnecessary stress or follow-up testing.

- False negatives could delay further clinical evaluation.
- Over reliance on the AI prediction by the user without clinical context could result in suboptimal decisions.
- Risk of data misuse if patient logs are not managed appropriately.

To mitigate these:

- Confidence scores are shown
- Users are informed the system is a decision support tool
- Logs are stored locally and securely

4. Human Oversight Measures (AI Act Article 14)

- Human professionals are responsible for all final decisions.
- SHAP explanations offer per-feature transparency to support informed interpretation.

The user interface is intentionally designed to:

- Require manual input or file upload by the clinician
- Clearly label outputs with human-readable conclusions
- Provide an “About” section and disclaimer to reinforce oversight

5. Input Specifications and Data Controls

- The system requires all 22 voice biomarkers to be present and formatted correctly.
- Input data is standardized using a pre-trained scaler to ensure consistency.
- Uploaded CSVs are validated for structure before prediction is allowed.

6. Alignment with ALTAI Principles

Human Agency Requires clinician oversight; no automation of clinical decisions

Technical Robustness Stable performance on known dataset; modular updates supported

Privacy No personal identifiers processed; logs stored locally

Transparency SHAP explanations integrated and visible in interface

Diversity & Fairness Not assessable due to lack of demographic labels in dataset

Societal Well-being Intended to improve early detection and reduce disease burden

Accountability Logs maintained per patient; user is informed of system boundaries

7. Performance Metrics and Suitability

The model was evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

Metrics are appropriate for the binary classification task with probabilistic output, and help balance sensitivity and specificity.

- SHAP is used to validate the logical consistency of individual predictions.

8. Lifecycle and Version Changes

- This prototype includes manual model and scaler versioning via .pkl files.
- No live update or retraining mechanism is included in the current version.

Future lifecycle improvements could include:

- Automated drift detection

- Logging review dashboards
- Support for model update triggers

9. Post-Market Monitoring and Performance Evaluation Plan

Logs are saved locally after each prediction, labeled by patient ID and timestamp. These logs can be reviewed by the deployer to:

- Analyze prediction frequency and consistency
- Identify misclassification or anomalous behavior
- Evaluate usage patterns across clinicians or contexts
- Although no centralized monitoring system is implemented, these local logs form a foundation for performance analysis in post-deployment phases.

Benefit and Risk Analysis

Identified risk	Description	Severity	Likelihood	Control measures	Residual risk
False negative	AI predicts no Parkinson's in affected patient	High	Medium	Human oversight, SHAP interpretation, warnings in UI	Moderate
False positive	AI predicts Parkinson's in healthy patient	Medium	Medium	Confidence score, follow up, clinician interpretation	Low
Input error	Incorrect data format or missing values	Medium	Medium	Input validation, form structure, column checking	Low
Misuse by unqualified user	Tool used outside of medical context	High	Low	Login system, disclaimers, warnings	Low
Model overfitting	Performance drops over time	Medium	Medium	Local logging, retraining capability	Moderate
Data privacy breach	Logs linked to identifiable patient data	High	Low	No identifiable information collected, local storage only	Very Low

3. Risk Management Report

Residual Risk Evaluation:

- All high-severity risks have been mitigated to an acceptable level given the context (prototyping, simulation, clinician use only).

- No unacceptable risks remain that would prevent safe use under intended conditions.

Benefit-Risk Justification:

- The potential benefits of early Parkinson's detection, improved clinician insight, and non-invasive testing outweigh the residual risks.

- Residual risks are further reduced by requiring the tool to be used in a controlled, supervised environment.

AI Lifecycle

Training Phase:

- Public dataset (UCI Parkinson's)
- Model and scaler serialized into version-controlled files

Deployment Phase:

- Local execution via Streamlit
- Model, scaler, and logs remain offline

Post-Deployment Monitoring:

- Manual log review via patient ID
- No automated retraining yet included

Upgrades:

- Modular design allows for periodic model retraining
- Logs can be used for error or bias analysis

Clinical Evaluation

Current State:

- No real patient testing yet conducted
- Internal validation on public dataset

Planned Enhancements:

- If developed further, external testing and formal accuracy analysis
- Collection of feedback from clinical users

Technical Documentation on Post-Market Surveillance

1. Post-Market Surveillance Objectives

- Monitor system performance and usage in a clinical setting (simulated)
- Detect possible model degradation, input anomalies, or user misuse

- Gather data to inform model retraining or revision
- Ensure continued alignment with safety, transparency, and ethical standards

2. Monitoring Methods

- Local logging: All predictions are logged by patient ID, timestamp, input values, model output, and confidence scores.
- Audit system: These logs can be reviewed by clinicians, researchers, or system administrators to assess:
  - Prediction consistency
  - Error trends (e.g. flagged misclassifications)
  - User input patterns (e.g. common mistakes or misuse)
- Log file format: Structured .csv logs ensure compatibility with spreadsheet tools and Python analytics libraries.

3. Periodic Review and Update Plan

The system does not include automated updates, a hypothetic approach would be:

- Monthly log audits (manual): Check for unusual prediction distributions or repeated feature anomalies
- Quarterly performance review: Re-evaluate performance using new or external datasets, if available
- Annual model revision (planned): Retrain the model if performance drifts or significant input changes are detected

4. Trigger-Based Monitoring Examples

- Performance degradation: A drop in prediction accuracy or confidence clustering may trigger a retraining request
- Feature drift: Unexpected trends in input distributions may indicate a need for model adaptation

- User feedback: If clinicians report errors or concerns, logs can be used to investigate and retrace model behavior

5. Future Considerations

- Integration of a central dashboard for reviewing logs and alerts
- Development of a feedback form to collect clinician experience data
- Exploration of federated learning or secure data contribution pipelines for privacy-preserving model updates

6. Compliance Goals

- Simulated PMS activities align with MDR Annex III and AI Act Articles 61 and 72
- Monitoring strategy supports transparency, user accountability, and safety across the device's use cycle

Disclaimer: This document simulates the design and compliance lifecycle of a Class IIa AI-based medical device intended to support clinical diagnosis of Parkinson's disease. It is developed as part of the Bachelor of Artificial Intelligence course "Laboratory of Medical Devices and Systems" and is not an approved or CE-marked medical product.