



+ Precision Medicine

VOICE-BASED DETECTION OF PARKINSON DISEASE

Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network

UC#17 – 2024/12 (v2.1)



PROBLEM DEFINITION

GOAL Design an AI-based decision system that accurately and instantly diagnoses Parkinson disease from simple voice recordings to obtain a rational diagnosis of the patient’s condition.

PROS & BENEFITS

- ▶ Identify the parameters involved in the Parkinson disease and enhance medical knowledge by helping neurologists understand the causal relationships between these parameters, their combination, and the disease.
- ▶ Help the medical profession to make earlier and more personalized decisions through rapid, systematic, and explainable diagnoses.
- ▶ Use a model with simple recordings to limit medical protocols that can be costly.

REFERENCE DATA

Source:
Sakar, C.O., Dept of Computer Engineering, Bahcesehir University, Istanbul, Serbes, G., Dept of Biomedical Engineering, Yildiz Technical University, Istanbul, Gunduz, A., Dept of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University-Cerrahpaşa, Nizam, H., Dept of Computer Engineering, Istanbul University-Cerrahpaşa, Sakar, B.E., Dept of Software Engineering, Bahcesehir University, Istanbul, Türkiye.

Dataset:
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

Variable to Predict: The model diagnoses the patient’s condition from the voice recordings as **PARKINSON | NO PARKINSON**

Potential Predictors: 753 potential predictors, clinically useful information from various speech signal processing applied on each recording [Time Frequency Features, Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform based Features, Vocal Fold Features and tunable Q-factor Wavelet transform (TQWT) features ...].

Observations: 756 reference voice recordings (for a total of 252 patients who sustained phonation of the vowel ‘a’ with 3 repetitions).

642 recordings compose a Learning Dataset for model induction using Training and Validation Datasets.

114 recordings compose an External Test Dataset to check the top-model’s performance on real unknown data and for benchmarking.

Learning Dataset: 642 recordings 84.92% 80% for Training, 20% for Validation		External Test Dataset: 114 recordings 15.08%	
NO PARKINSON	PARKINSON	NO PARKINSON	PARKINSON
163 25.4%	479 74.6%	29 25.4%	85 74.6%

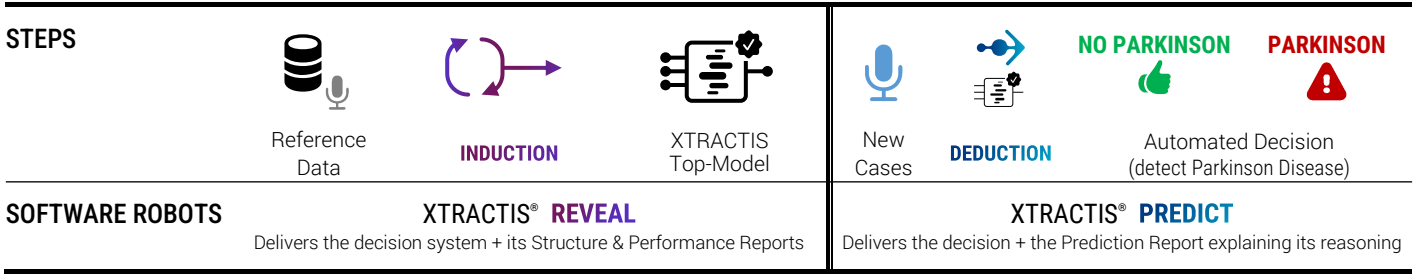
MODEL TYPE

Regression Multinomial Classification **Binomial Classification** Scoring

XTRACTIS-INDUCED DECISION SYSTEM

- Intelligible Model, Explainable Decisions**
 - ▶ The top-model is a decision system composed of 26 gradual rules without chaining aggregated into 2 disjunctive rules.
 - ▶ Each rule uses from 1 to 24 predictors among the 92 variables that XTRACTIS automatically identified as significant (out of the 753 features describing each recording).
 - ▶ Only a few rules are triggered at a time to compute the decision.
- High Predictive Capacity** It has a pretty good Real Performance (on unknown data).
- Ready to Deploy** It computes real-time predictions up to 70,000 decisions/second, offline or online (API).

XTRACTIS PROCESS



TOP-MODEL INDUCTION

INDUCTION PARAMETERS & PROCESS

Powered by:



- We launch 1,000 inductive reasoning strategies. Due to the small number of reference cases, each strategy is applied to 20 different 5-fold-partitions of the Learning Dataset to get a reliable assessment of the descriptive and predictive performances, respectively from Training and Validation Datasets.
- Each strategy thus generates 100 unitary models called **Individual Virtual Expert (IVE)**, whose decisions are aggregated with 3 possible operators into a **College of Virtual Experts (CVE)**.
- Among the 3,000 induced CVEs, the top-CVE with the best predictive performance remains complex: 2,014 rules share 464 predictors.

Given the small number of cases in the reference dataset, the XTRACTIS **CVE→IVE** Reverse-Engineering process is necessary to induce a unitary intelligible model through a single split cross-validation, from a large synthetic reference dataset:

- We build a synthetic dataset composed of 32,100 new cases simulated by deduction from the top-CVE, around the 642 original learning cases but distinct from them.
- We apply 500 induction strategies to the same single partition of this new dataset (34% Training | 33% Validation | 33% Test): XTRACTIS induces 500 IVEs.
- The top-IVE selected is the one that has the best performance and with the best intelligibility, i.e., the fewer predictors and rules.

Total number of induced unitary models

100,500 IVEs

Criterion for the induction optimization

F1-score

Validation criterion for the top-model selection

F1-score

Duration of the process @ Induction Speed FP64

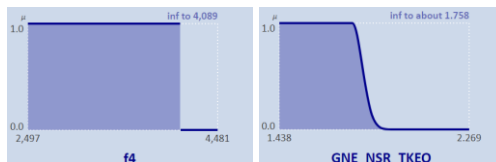
75.8 days @ 1.00 Tflops

TOP-MODEL STRUCTURE

The top-IVE has a rather poor intelligibility as it has **26 rules** combining **92 predictors**, with 9.3 predictors per rule on average. But it remains acceptable given the high level of complexity of the phenomenon having initially 753 potential predictors. This model's Structure Report reveals all the internal logic of the decision system and ensures that the model is understandable. It is a transparent model that can be audited by the expert and certified by the regulator before deployment to end-users.

PREDICTORS

- 92 features identified out of 753
- Ranked by individual contribution (1 strong, 20 medium & 70 weak signals): #1 `tqwt_maxValue_dec_19` / #2 `tqwt_stdValue_dec_5` / ...
- Labeled by nominal, binary, and fuzzy classes
 Examples: **binary nominal** "{Male}"
binary interval "inferior to 4,089"
fuzzy interval "inferior to about 1.758"



RULES

- 26 connective fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules)
- 1 to 24 predictors per rule (on average, 9.3 predictors per rule)
- Example: fuzzy rule **R8** uses 6 predictors and concludes **PARKINSON**. 25 other rules complete this model.

```

IF gender IS {Male}
AND f4 IS inferior to 4,089
AND GNE_NSR_TKEO IS inferior to ~1.758
AND tqwt_energy_dec_18 IS inferior to ~0.206
AND tqwt_minValue_dec_9 IS superior to ~-0.142
AND tqwt_maxValue_dec_26 IS inferior to ~-1.12
THEN Diagnosis IS PARKINSON
    
```

TOP-MODEL PERFORMANCE

The top-IVE performances, measured in Training/Validation/Test on synthetic data, and on original points, then in External Test on reference data, guarantee the model's predictive and real performances.

Perf. Type	Quality of CVE copy			642 original points
	34% Training (Synthetic Data)	33% Validation (Synthetic Data)	33% Test (Synthetic Data)	
F1-score	97.73%	97.84%	97.18%	95.09%
Classification Error	1.14%	1.08%	1.42%	2.49%

REAL
External Test
88.46%
5.45%

EXPLAINED PREDICTIONS FOR 3 UNKNOWN CASES

CASE

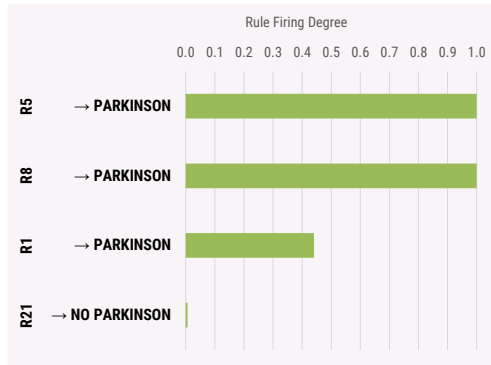
(from the External Dataset, i.e., not included in the Learning Dataset)

RECORDING #217

actual value = PARKINSON	
PPE	0.757
DFA*	0.853
...	...
gender	Male

DEDUCTIVE INFERENCE OF RULES

For this case, 4 rules are triggered:
R5 and **R8** are fired at 1.000, and **R1** at 0.441 to conclude PARKINSON,
R21 at 0.006 to conclude NO PARKINSON.
 22 other rules are not activated.



AUTOMATED DECISION

NUMBER OF TRIGGERED RULES
4 / 26

FUZZY PREDICTION
{ PARKINSON | 1.000,
NO PARKINSON | 0.006 }

FINAL PREDICTION
{ PARKINSON }

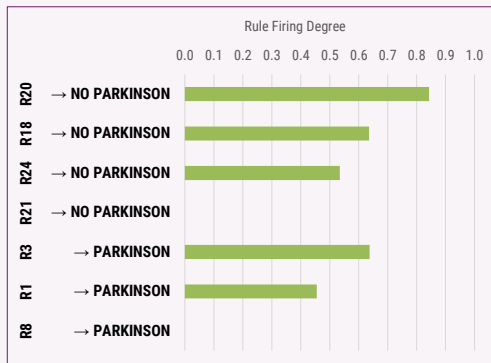
The system delivers a correct detection compared to the observed case:

PARKINSON 

RECORDING #34

actual value = NO PARKINSON	
PPE	0.821
DFA	0.668
...	...
gender	Male

For this case, 7 rules are triggered:
R20 is fired at 0.843, **R18** at 0.636, **R24** at 0.535 and **R21** at 2.16e-04 to conclude NO PARKINSON,
R3 at 0.445, **R1** at 0.455 and **R8** at 6.32e-08 to conclude PARKINSON.
 19 other rules are not activated.



NUMBER OF TRIGGERED RULES
7 / 26

FUZZY PREDICTION
{ NO PARKINSON | 0.843,
PARKINSON | 0.638 }

FINAL PREDICTION
{ NO PARKINSON }

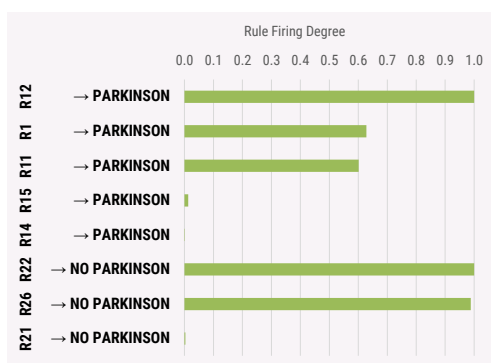
The system delivers a correct detection compared to the observed case:

NO PARKINSON 

RECORDING #53

actual value = PARKINSON	
PPE	0.834
DFA	0.579
...	...
gender	Female

For this case, 8 rules are triggered:
R12 is fired at 1.000, **R1** at 0.628, **R11** at 0.601, **R15** at 0.013 and **R14** at 0.001 to conclude PARKINSON,
R22 at 1.000, **R26** at 0.988 and **R21** at 0.003 to conclude NO PARKINSON.
 16 other rules are not activated.



NUMBER OF TRIGGERED RULES
8 / 26

FUZZY PREDICTION
{ PARKINSON | 1.000,
NO PARKINSON | 1.000 }


FINAL PREDICTION
REFUSAL

The system cannot decide between the 2 classes, so it refuses to decide.

More reference data near this recording profile should eliminate the undecidability of the updated model in this decision space area.

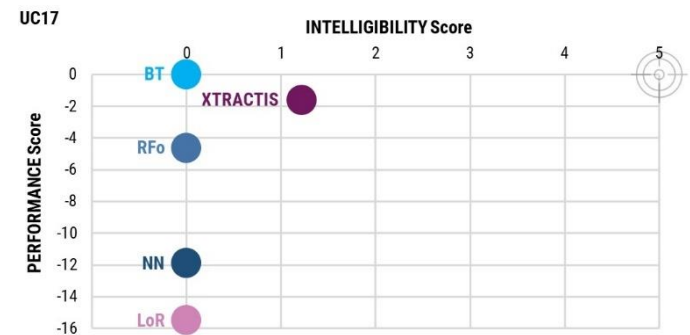
*Predictor value outside the variation range of the model but inside the allowed extrapolation range. XTRACTIS will refuse to give a result for an extrapolation far from the allowed extrapolation range. It is one situation of the "Refusal" prediction.

TOP-MODELS BENCHMARK

	XTRACTIS 	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK	
MODELING PARAMETERS	MODELS RELEASE	2023/05	2023/05	2023/05	2023/05	
	ALGORITHM VERSION	XTRACTIS REVEAL 13.0.45667	XTRACTIS BENCHMARK module embedding Python 3.9.10 Scikit-Learn 1.1.2 LightGBM 3.3.2 TensorFlow 2.10.0 Keras 2.10.0			
	CROSS-VALIDATION TECHNIQUE	20x5 folds for each CVE model. Then 1-Split Validation for each IVE model: 34% Training 33% Validation 33% Test	20x5 folds for each CVE model			
	NUMBER OF EXPLORED STRATEGIES⁽¹⁾	1,000 induction strategies for the CVE on Training / Validation data. 3 aggregation operators tested. 500 induction strategies for the IVE on synthetic data	1,000 ML strategies on Training / Validation data. Aggregation operator: Absolute Majority.			
	TOP-MODEL SELECTION⁽²⁾	Top-CVE with Relative Majority Aggregator selected among the 3,000 CVEs. Then Top-IVE among 500 IVEs	Top-CVE selected among 1,000 CVEs, then single model obtained by applying best CVE strategy on 100% of the Learning Dataset			

TOP-MODEL STRUCTURE	NUMBER OF PREDICTORS (out of 753 Potential Predictors)	92	610	366	431	753
	AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	9.3 per rule	610.0 per equation	7.5 per rule	4.7 per rule	254.6 per equation
	STRUCTURE OF THE DECISION SYSTEM	26 fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules) Only a few rules are triggered at a time to compute a prediction	1 linear equation	56 trees without chaining 2,416 binary rules	1 chain of 110 trees 1,467 binary rules Tree #N corrects the error of the N-1 previous trees	3 hidden layers 42 hidden nodes 43 equations 42 unintelligible synthetic variables, in addition to the 753 original predictors

TOP-MODEL SCORES		Random ⁽³⁾	XTRACTIS	LoR	RFo	BT	NN
	INTELLIGIBILITY Score⁽⁴⁾		1.22	0.00	0.00	0.00	0.00
	CVE Real Performance (F ₁ -Score) in External Test	51.72	86.79	74.07	86.21	87.27	81.63
	Gap to CVE Leader in External Test	-35,55	-0.48	-13.20	-1.06	0.00	-5.64
	IVE Real Performance (F ₁ -Score) in External Test	51.72	88.46	73.47	83.02	91.23	73.08
	Gap to IVE Leader in External Test	-39,51	-2.77	-17.76	-8.21	0.00	-18.15
Average Real Performance in External Test	51.72	87.63	73.77	84.62	89.25	77.36	
PERFORMANCE Score⁽⁴⁾			-1.63	-15.48	-4.64	0.00	-11.90



(1) For all algos: on exactly the same splits of the Learning Dataset. All Models are optimized according to their Validation F₁-Score.

(2) All top-models are selected according to their Validation F₁-Score while checking that it remains close to their Training F₁-Score.

(3) Baseline performances that models must exceed to perform better than chance (P-value = 0.001; 100,000 models generated by random permutation of the output values). The value of each performance criterion is generally achieved by a different random model.

(4) See Appendices for explanations and detailed results. Performance Scores are calculated on all available unknown data.

More Use Cases:
xtractis.ai/use-cases/

APPENDIX 1 – Calculation of the Intelligibility × Performance Scores

AI Technique #i	T _i	i ∈ [1 ; n] n = number of AI Techniques benchmarked in terms of data-driven modeling = 5
Benchmark #k	B _k	k ∈ [1 ; p] p = number of Benchmarks for the Use Case ∈ {1, 2, 3}

Remarks:

- In case of a small number of reference data, a CVE model (College of Virtual Experts) is generated by each explored strategy of T_i, generally via an N×K-fold cross validation. In this case, a Benchmark is led with the top-CVE on the External Test Dataset (ETD, composed of unknown reference cases). Then, a top-IVE model (Individual Virtual Expert) is generated from the top-CVE, through the XTRACTIS® reverse-engineering process, or for the other T_i, by applying the top-strategy, which has generated the top-CVE, on the Training and Validation Datasets. And a second Benchmark is led with this top-IVE on the same ETD.
- In case of a huge number of reference data, an IVE is generated by each explored strategy of T_i, via a 1-split validation. In this case, Benchmarks are led with the top-IVE on the Test Dataset (TD, composed of unknown reference cases) and on the available ETDs.
- Each Benchmark uses the latest versions of the following algorithms available at the date of the benchmark. XTRACTIS®: REVEAL; Logistic Regression: Python, Scikit-Learn; Random Forest & Boosted Tree: Python, LightGBM; Neural Network: Python, TensorFlow, Keras.
- Each B_k uses exactly the same TD and ETD for each T_i model.
- No Regression models can be obtained by Logistic Regression. So, this Data Analysis technique is benchmarked only for Classification or Scoring problems.
- The Holy Grail for critical AI-based decision systems is to obtain a model with the highest Performance and the highest Intelligibility scores (top-right corner of the graph).

PERFORMANCE Score

For each B_k, we calculate the values of the Performance Criterion (PC) on the same ETD for all the T_i top-CVEs; and on the same TD and ETDs for all the T_i top-IVEs. The PC is: RMSE in percentage for a Regression; F₁-Score for a Binomial Classification; Average F₁-Score or Average F₂-Score for a Multinomial Classification; Gini index for a Scoring. Then, we compare the value of the PC of each T_i top-CVE (resp. top-IVE) to the best value of this PC reached by the best T_i top-CVE (resp. top-IVE) on ETD (resp. on TD and ETDs).

For Regression, we calculate for each T_i top-model (CVE and IVE): PS(T_i, B_k) = Best_PC(B_k) - PC(T_i, B_k).

For Classification and Scoring, we calculate for each T_i top-model: PS(T_i, B_k) = PC(T_i, B_k) - Best_PC(B_k).

$$\text{Performance Score of } T_i$$

$$\text{PS}(T_i) = \text{Mean} (\text{PS}(T_i, B_k))_{k \in [1 ; p]}$$

Remark:

- Each PS varies theoretically from -100 (Lowest Score) to 0 (Highest Score), but practically between -50 and 0.

INTELLIGIBILITY Score

We consider the T_i top-IVE. Its Intelligibility Score IS(T_i) is valued from 0.00 to 5.00 regarding the structure of the model: number of predictors, classes, rules, equations, trees, synthetic variables, modalities to predict for classifications (or numeric variables to predict for regressions or scoring). The more compact the model, the higher its IS.

The IS of each T_i is obtained by accumulating the following five penalty values to the ideal IS value of 5.00 (each penalty has a null or a negative value):

- Penalty 1 (logarithmic penalty regarding the number of predictors):

$$\text{Pen1}(T_i) = \min(0, 1 - \log_{10} \text{number of predictors})$$

Examples: Pen1 = 0.00 for up to 10 predictors
Pen1 = -3.00 for 10.000 predictors

- Penalty 2 (linear penalty regarding the average number of rules or equations per modality to predict):

$$\text{Pen2}(T_i) = \min\left(0, 0.01 - \frac{\text{average number of rules or equations per modality to predict}}{100}\right)$$

Examples: Pen2 = 0.00 for 1 rule or equation per modality to predict on average
Pen2 = -3.00 for 301 rules or equations per modality to predict on average

- Penalty 3 (linear penalty regarding the average number of predictors per rule or equation):

$$\text{Pen3}(T_i) = \min\left(0, \frac{9 - 3 \times \text{average number of predictors per rule or equation}}{7}\right)$$

Examples: Pen3 = 0.00 for up to 3.0 predictors per rule or equation on average
Pen3 = -3.00 for 10.0 predictors per rule or equation on average

- Penalty 4 (linear penalty regarding the number of chained trees, here for BT only):

$$\text{Pen4}(T_i) = \min(0, 1 - \text{number of chained trees})$$

Examples: Pen4 = 0.00 for 1 tree
Pen4 = -3.00 for 4 chained trees

- Penalty 5 (maximum penalty due to unintelligibility of synthetic variables, here for NN only):

$$\text{Pen5}(T_i) = -5$$

Intelligibility Score of T_i

$$\text{IS}(T_i) = \max(0.00, 5.00 + (\text{Pen1} + \text{Pen2} + \text{Pen3} + \text{Pen4} + \text{Pen5}))$$

Remarks:

- For the difference between the Intelligibility and the Explainability of a model, please see the XTRACTIS® Brochure, page 7.
- The real complexity of the process/phenomenon under study is intrinsic, i.e., it could not be reduced or simplified, but only discovered; thus, the top-model will be complex if the process/phenomenon turns out to be complex [Zalila 2017]. Consequently, for some complex process/phenomenon, IS can be equal to 3.00 or less, even if T_i natively produces intelligible models (XTRACTIS, Random Forest).
- For similar structures, the Boosted Tree model is always less intelligible than the Random Forest one, as it is composed of chains of trees, instead of a college of trees (see Penalty 4).
- Neural Network model has always the lowest IS of 0.00, because it uses synthetic unintelligible variables (hidden nodes) in addition to all the potential predictors (see Penalty 5).

APPENDIX 2 – Use Case Results (all Performance criteria of all Top-Models)

Performance Criterion	Classification Error	Min. Sensitivity Specificity	Sensitivity	Specificity	PPV	NPV	F ₁ -Score	Refusal
-----------------------	----------------------	------------------------------	-------------	-------------	-----	-----	-----------------------	---------

RANDOM MODEL

Nb of Random Permutations (P-value) = 100,000 (0.001%)

Performance against chance (External Test)	24.56%	51.72%					51.72%	
--	--------	--------	--	--	--	--	--------	--

XTRACTIS TOP-MODEL

CVE - Descriptive Performance (Training)	1.09%	98.75%	99.39%	98.75%	96.43%	99.79%	97.89%	0 (0.00%)
CVE - Predictive Performance (Validation)	1.25%	96.32%	96.32%	99.58%	98.74%	98.75%	97.52%	2 (0.31%)
CVE - Real Performance (External Test)	6.14%	79.31%	79.31%	98.82%	95.83%	93.33%	86.79%	0 (0.00%)
IVE - Descriptive Performance (Training)	1.14%	97.66%	97.66%	99.27%	97.81%	99.22%	97.73%	7 (0.06%)
IVE - Predictive Performance (Validation)	1.08%	97.37%	97.37%	99.45%	98.33%	99.12%	97.84%	5 (0.05%)
IVE - Real Performance (Test)	1.42%	97.18%	97.18%	99.05%	97.18%	99.05%	97.18%	12 (0.11%)
IVE - Real Performance (642 original points)	2.49%	95.09%	95.09%	98.33%	95.09%	98.33%	95.09%	0 (0.00%)
IVE - Real Performance (External Test)	5.45%	82.14%	82.14%	98.78%	95.83%	94.19%	88.46%	4 (3.51%)

LOGISTIC REGRESSION TOP-MODEL

CVE - Descriptive Performance (Training)	6.39%	87.73%	87.73%	95.62%	87.20%	95.82%	87.46%	
CVE - Predictive Performance (Validation)	12.31%	73.62%	73.62%	92.48%	76.92%	91.15%	75.24%	
CVE - Real Performance (External Test)	12.28%	68.97%	68.97%	94.12%	80.00%	89.89%	74.07%	
IVE - Descriptive Performance (Training)	6.39%	84.66%	84.66%	96.66%	89.61%	94.88%	87.07%	
IVE - Real Performance (External Test)	11.40%	62.07%	62.07%	97.65%	90.00%	88.30%	73.47%	

RANDOM FOREST TOP-MODEL

CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Predictive Performance (Validation)	5.45%	86.50%	86.50%	97.29%	91.56%	95.49%	88.96%	
CVE - Real Performance (External Test)	7.02%	86.21%	86.21%	95.29%	86.21%	95.29%	86.21%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	7.89%	75.86%	75.86%	97.65%	91.67%	92.22%	83.02%	

BOOSTED TREE TOP-MODEL

CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Predictive Performance (Validation)	4.67%	88.34%	88.34%	97.70%	92.90%	96.10%	90.57%	
CVE - Real Performance (External Test)	6.14%	82.76%	82.76%	97.65%	92.31%	94.32%	87.27%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	4.39%	89.66%	89.66%	97.65%	92.86%	96.51%	91.23%	

NEURAL NETWORK TOP-MODEL

CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Predictive Performance (Validation)	4.67%	87.73%	87.73%	97.91%	93.46%	95.91%	90.51%	
CVE - Real Performance (External Test)	7.89%	68.97%	68.97%	100.00%	100.00%	90.43%	81.63%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	12.28%	65.52%	65.52%	95.29%	82.61%	89.01%	73.08%	

The entirety of this document is protected by copyright. All rights are reserved, particularly the rights of reproduction and distribution. Quotations from any part of the document must necessarily include the following reference:
Zalila, Z., Intellictech & Xtractis (2023-2024). XTRACTIS® the General Reasoning AI for Trusted Decisions. Use Case #17 | Precision Medicine: Voice-based Detection of Parkinson Disease – Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network. INTELLITECH [intelligent technologies], December 2024, v2.1, Compiègne, France, 6p.