



+ Precision Medicine

SPECTROMETRIC DIAGNOSIS OF OVARIAN CANCER

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

UC#05 – 2025/06 (v3.0)

xtractis.ai

PROBLEM DEFINITION

GOAL Design an AI-based decision system that accurately and instantly makes a rational medical diagnosis of ovarian cancer, from a sample of serum analyzed by a mass spectrometer.

PROS & BENEFITS

- ▶ Identify the proteins involved in cancer, from the spectrum bands.
- ▶ Enhance medical knowledge by helping gynecologists and oncologists understand the causal relationships between specific proteins, their combination, and the presence of cancer.
- ▶ Help the medical profession to make earlier and more personalized decisions through rapid, systematic, and explainable diagnoses.
- ▶ Contribute to improving patient care (pain, survival, duration of treatment) and extend access to high-level diagnoses even in medical deserts.

REFERENCE DATA

Source: NCI PBSII, Emanuel F. Petricoin and al, Food and Drug Administration / National Institutes of Health Clinical Proteomics Program, Department of Therapeutic Proteins/Center for Biologics Evaluation and Research, FDA, Bethesda, MD, USA.

Dataset: <https://leu.ugr.es/elvira/DBCRepository/OvarianCancer/OvarianCancer-NCI-PBSII.html>

Variable to Predict: The model makes the diagnosis of the sample serum as **TUMOR | NORMAL**.

Potential Predictors: 15,154 parameters are m/z ratios (mass/charge) originating from the spectrum of each sample.

Observations: 253 samples of serum from patients with or without ovarian cancer, characterized by their mass spectrum.

Data are divided into a Learning Dataset for model induction using Training and Validation Datasets, and an External Test Dataset to check the top-model's performance on real data and for benchmarking.

Learning Dataset: : 169 66.8% cases 80% for Training, 20% for Validation	
NORMAL	TUMOR
108 63.9%	61 36.1%

External Test Dataset: 84 33.2% cases	
NORMAL	TUMOR
54 64.3%	30 35.7%

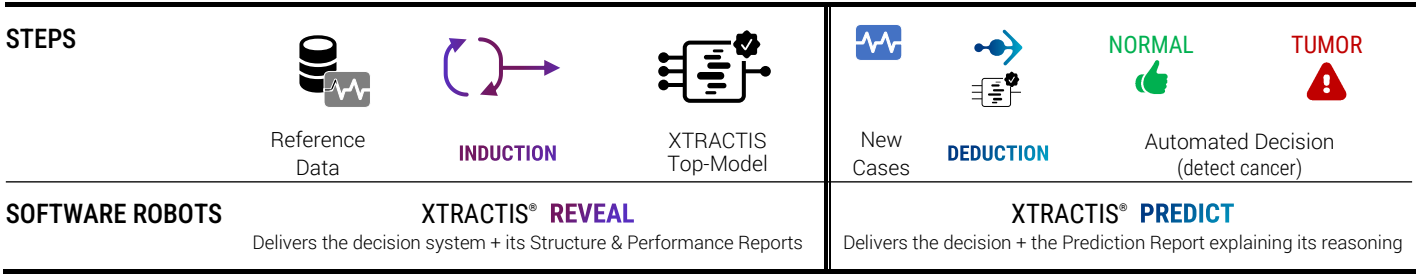
MODEL TYPE

Regression Multinomial Classification **Binomial Classification** Scoring

XTRACTIS-INDUCED DECISION SYSTEM

- Intelligible Model, Explainable Decisions**
 - ▶ The top-model is a decision system composed of 2 disjunctive gradual rules without chaining. *Remark: Even if the theoretical complexity of this problem was very high, the decision process studied turns out to be quite simple, although non-linear.*
 - ▶ Each rule uses from 1 to 3 predictors among the 3 variables that XTRACTIS automatically identified as significant (out of the 15,154 level of genes expression describing each patient).
 - ▶ Rules are not necessarily triggered at the same time to compute the decision.
- High Predictive Capacity** It has a perfect Real Performance (on unknown data).
- Efficient AI System** It computes real-time predictions up to 70,000 decisions/second, offline or online (API).

XTRACTIS PROCESS



TOP-MODEL INDUCTION

INDUCTION PARAMETERS

Powered by:



- We launch 300 inductive reasoning strategies; each strategy is applied to 40 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 200 unitary models called **Individual Virtual Expert (IVE)**, whose decisions are aggregated with 3 possible operators into a **College of Virtual Experts (CVE)**.
- Among the 900 induced CVEs, the top-CVE with the best predictive performance remains complex: 555 rules sharing 151 predictors.

Given the small number of reference cases in the reference dataset, the XTRACTIS **CVE→IVE** Reverse-Engineering process is necessary to get a more intelligible model:

- We build a synthetic dataset composed of 20,280 new cases simulated by deduction from the top-CVE, around the 169 original learning cases but distinct from them.
- We apply 300 induction strategies to the same single 34% Training | 33% Validation | 33% Test partition of this new dataset: XTRACTIS induces 300 IVEs.
- The top-IVE selected is the one that is the most intelligible while being as efficient as the top-CVE.

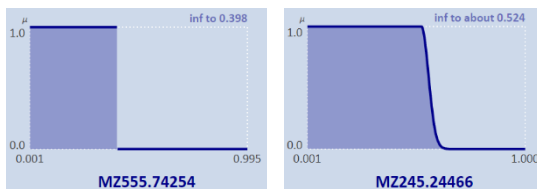
Total number of induced unitary models	Criterion for the induction optimization	Validation criterion for the top-model selection	Duration of the process (Induction Power FP64)
60,300 IVEs	F₁-Score	F₁-Score	41 days & 18 hours (1.13 Tflops)

TOP-MODEL STRUCTURE

The top-IVE model has an excellent intelligibility -and is very simple- as it combines the 3 predictors into only 2 rules with 2 predictor per rule on average. Its Structure Report reveals all the internal logic of the decision system and ensures that the model is understandable by the human expert. It is a transparent model that can be audited and certified before deployment to end-users.

PREDICTORS

- 3 spectrum m/z ratios out of 15,154
- Ranked by individual contribution (1 strong, 1 medium & 1 weak signal):
#1 [MZ2.8234234](#) / #2 [MZ245.24466](#) / #3 [MZ555.74254](#)
- Labeled by fuzzy and binary classes
Examples: **binary interval** "inf to 0.398";
fuzzy interval "inf to about 0.524"



RULES

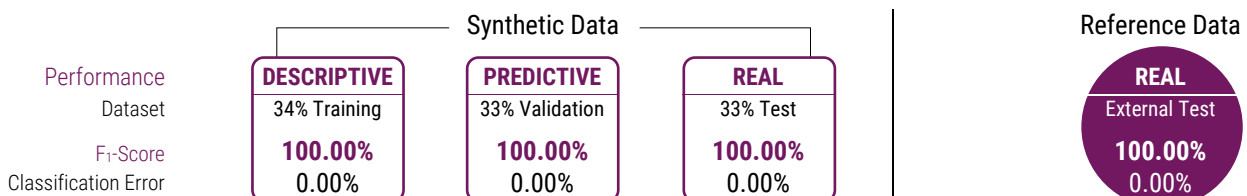
- 2 connective fuzzy rules without chaining
- 1 to 3 predictors per rule (on average, 2 predictors per rule)
- Example: fuzzy rule **R2** uses 3 predictors and concludes "TUMOR". Fuzzy rule R1 completes this model.

IF	MZ2.8234234	IS	inf to 0.523
AND	MZ245.24466	IS	inf to ~0.524
AND	MZ555.74254	IS	inf to 0.398
THEN	Diagnosis	IS	TUMOR

Literally, the sampled serum gets a tumor diagnosis if the [MZ2.8234234](#) spectrum ratio is inferior to 0.523, and the [MZ245.24466](#) spectrum ratio is under around 0.524, and the [MZ555.74254](#) spectrum ratio is inferior to 0.398.

TOP-MODEL PERFORMANCE

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



EXPLAINED PREDICTIONS FOR 2 UNKNOWN CASES

CASE

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

DEDUCTIVE INFERENCE OF RULES

AUTOMATED DECISION

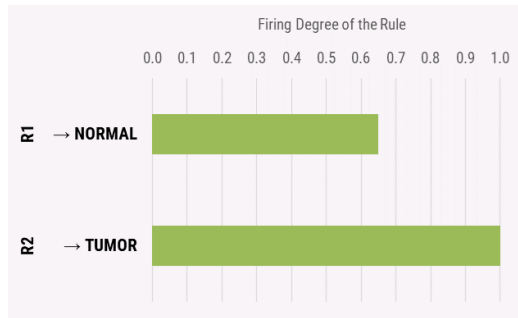
PATIENT #97

actual value = TUMOR

MZ2.8234234	0.332
MZ245.24466	0.099
MZ555.74254	0.163



For this patient, the 2 rules are triggered:
R2 is fired at 1.000, and R1 at 0.648.



NUMBER OF TRIGGERED RULES

2 / 2

FUZZY PREDICTION

{ TUMOR | 1.000, NORMAL | 0.648 }

FINAL PREDICTION

{ TUMOR }

The system delivers a correct diagnosis of cancer compared to that given by the oncologist:

TUMOR



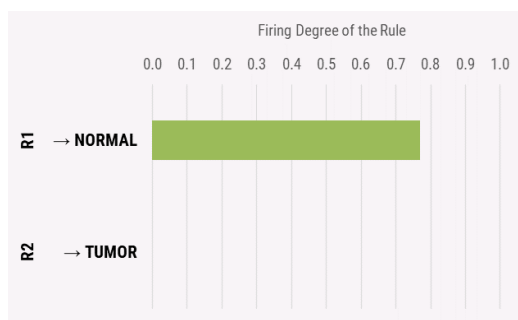
PATIENT #16

actual value = NORMAL

MZ2.8234234	0.374
MZ245.24466	0.808
MZ555.74254	0.138



For this patient, only 1 rule is triggered:
R1 is fired at 0.769.
R2 is not activated.



NUMBER OF TRIGGERED RULES

1 / 2

FUZZY PREDICTION

{ NORMAL | 0.769 }

FINAL PREDICTION


{ NORMAL }

The system delivers a correct diagnosis of cancer compared to that given by the oncologist:

NORMAL

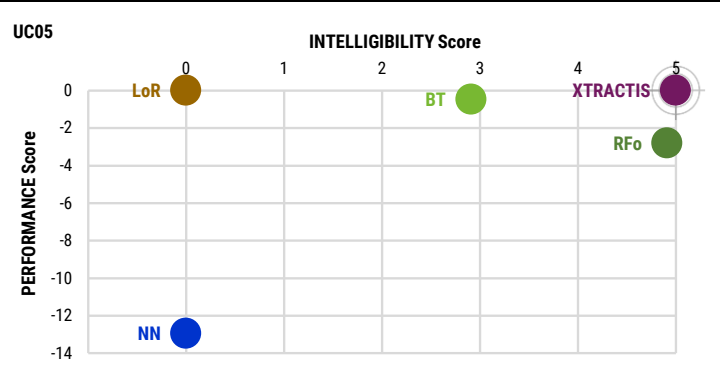


TOP-MODELS BENCHMARK: DECISION STRUCTURE & INTELLIGIBILITY × PERFORMANCE SCORES

	XTRACTIS 	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREES	NEURAL NETWORK	
MODELING PARAMETERS	MODELS RELEASE	2022/06	2022/09	2022/08	2022/07	2022/09
	ALGORITHM VERSION	XTRACTIS REVEAL 12.1.42004	Python 3.7; Scikit-learn 1.0.2	Python 3.6; LightGBM 2.2.2	Python 3.6; LightGBM 2.2.2	Python 3.9; TensorFlow 2.9.1, Keras 2.9.0
	CROSS-VALIDATION TECHNIQUE	40x5 folds for each CVE model. Then 1-Split Validation for each IVE model: 34% Training 33% Validation 33% Test	40x5 folds for each CVE model	40x5 folds for each CVE model	40x5 folds for each CVE model	40x5 folds for each CVE model
	NUMBER OF EXPLORED STRATEGIES⁽¹⁾	300 induction strategies for the CVE on Training / Validation data. 300 induction strategies for the IVE on synthetic data	300 data analysis strategies on Training / Validation data	300 ML strategies on Training / Validation data	300 ML strategies on Training / Validation data	300 ML strategies on Training / Validation data
	TOP-MODEL SELECTION⁽²⁾	Top-CVE among 900 CVEs. Then Top-IVE among 300 IVEs	Top-CVE selected among 300 CVEs, then single model obtained by applying best CVE strategy on 100% of the Learning Dataset			

TOP-MODEL STRUCTURE	NUMBER OF PREDICTORS (out of 15,154 Potential Predictors)	3	35	4	6	15,154
	AVERAGE NUMBER OF PREDICTORS PER RULE / EQUATION	2.0 per rule	35 per equation	1.7 per rule	1.7 per rule	6496.3 per equation
	STRUCTURE OF THE DECISION SYSTEM	2 fuzzy rules without chaining	1 linear equation	3 trees 9 binary rules	1 chain of 3 trees 9 binary rules Tree #N corrects the error of the N-1 previous trees	2 hidden layers 6 hidden nodes 7 equations 6 unintelligible synthetic variables

TOP-MODEL SCORES		Random ⁽³⁾	XTRACTIS	LoR	RFo	BT	NN
	INTELLIGIBILITY Score⁽⁴⁾		5.00	0.00	4.91	2.91	0.00
	CVE Real Performance (F ₁ -Score) in External Test		100.00	100.00	97.20	100.00	95.41
	Gap to CVE Leader in External Test		0.00	0.00	-2.80	0.00	-4.59
	IVE Real Performance (F ₁ -Score) in External Test	77.78%	100.00	100.00	97.20	99.08	78.72
Gap to IVE Leader in External Test		0.00	0.00	-2.80	-0.92	-21.28	
Average Real Performance in External Test	77.78%	100.00	100.00	97.20	99.54	87.07	
PERFORMANCE Score⁽⁴⁾		0.00	0.00	-2.80	-0.46	-12.94	



(1) For all algos: on the same 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. Perfect results of XTRACTIS and LoR on External Test could be explained by a low number of reference points compared to the very large number of potential predictors.

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 variable or modality to predict):

$$\text{Pen2}(T_i) = \min\left(0, \frac{1 - \text{average number of rules or equations per variable or modality to predict}}{40}\right)$$

Examples: Pen2 = 0.00 for 1 rule or equation per variable or modality to predict on average
Pen2 = -3.00 for 121 rules or equations per variable or 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	28.57%	60.00%					77.78%	
----------------------------	--------	--------	--	--	--	--	--------	--

XTRACTIS TOP-MODEL

CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
CVE - Predictive Performance (Validation)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
CVE - Real Performance (External Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
IVE - Predictive Performance (Validation)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
IVE - Real Performance (Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
IVE - Real Performance (169 original points)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
IVE - Real Performance (External Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)

LOGISTIC REGRESSION TOP-MODEL

CVE - Descriptive Performance (Training)	0.59%	98.36%	100.00%	98.36%	99.08%	100.00%	99.54%	
CVE - Predictive Performance (Validation)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Real Performance (External Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

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)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Real Performance (External Test)	3.57%	96.30%	96.30%	96.67%	98.11%	93.55%	97.20%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	3.57%	96.30%	96.30%	96.67%	98.11%	93.55%	97.20%	

BOOSTED TREES TOP-MODEL

CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Predictive Performance (Validation)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Real Performance (External Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Descriptive Performance (Training)	0.59%	98.36%	100.00%	98.36%	99.08%	100.00%	99.54%	
IVE - Real Performance (External Test)	1.19%	96.67%	100.00%	96.67%	98.18%	100.00%	99.08%	

NEURAL NETWORK TOP-MODEL

CVE - Descriptive Performance (Training)	1.77%	98.15%	98.15%	98.36%	99.07%	96.77%	98.60%	
CVE - Predictive Performance (Validation)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Real Performance (External Test)	5.95%	90.00%	96.30%	90.00%	94.55%	93.10%	95.41%	
IVE - Descriptive Performance (Training)	16.57%	75.00%	75.00%	98.36%	98.78%	68.97%	85.26%	
IVE - Real Performance (External Test)	23.81%	68.52%	68.52%	90.00%	92.50%	61.36%	78.72%	

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., Idagrai Labs & Xtractis (2018-2025). XTRACTIS® the Reasoning AI for Trusted Decisions. Use Case #05 | Precision Medicine: Spectrometric Diagnosis of Ovarian Cancer – Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network. IDAGRAI LABS, June 2025, v3.0, Compiègne, France, 6p.