



Naval Security

IDENTIFICATION OF UNDERWATER SOUNDS (VIRTUAL GOLDEN EAR)

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

UC#29 – 2025/06 (v2.0)



PROBLEM DEFINITION

GOAL Design an AI-based decision-making system that accurately and rationally identifies underwater sounds from their signal characteristics.

PROS & BENEFITS

- ▶ Identify the specific parameters involved in the identification of underwater sounds and enhance knowledge by helping submarine staff and acoustic experts understand the causal relationships between these parameters, their combination, and the type of sound.
- ▶ Help to design a virtual "Golden Ear" (expert in underwater acoustics) operating 24/7/365 with the same quality of decision, or to design by simulation undetectable objects, or to design a 24/7/365 tutor for Human Golden Ear apprentices.
- ▶ Assist the military profession in making more reliable and quicker decisions, thanks to rapid, systematic, and explainable identification process with usual sensors.
- ▶ Avoid many false alarms thanks to transparent and accurate diagnosis.

REFERENCE DATA

Source:
EVIDEN, 2024:
initial data from NOAA NCEI (National Oceanic & Atmospheric Administration / National Centers for Environment Information) & NPS (National Park service)

Dataset:
Intelligible Preprocessing by EVIDEN-BDS expert (2024/06) from:
www.ncei.noaa.gov/products/passive-acoustic-data

Variable to Predict The model predicts the Sound Type among 5 modalities:
sea noise | click | propulsion | unclassified | vocalize

Potential Predictors **23 features characterize each sound signal** (21 are numeric variables and 2 are nominal variables): [step, fit, iband, band_freq_ratio, band_time_ratio, nharmonic, max_freq, max_ontime1, nreliance, time_base2...]. Original data were preprocessed in order to obtain intelligible features, understandable by experts.

Observations **549 reference recordings.**
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*: 464 cases 84.52%					External Test Dataset**: 85 cases 15.48%				
Training (371 80%), Validation (93 20%)									
sea noise	click	propulsion	unclassified	vocalize	sea noise	click	propulsion	unclassified	vocalize
69 14.87%	29 6.25%	45 9.70%	303 65.30%	18 3.88%	13 15.29%	6 7.06%	8 9.41%	54 63.53%	4 4.71%
*240 missing values (2.25%)					**47 missing values (2.40%)				

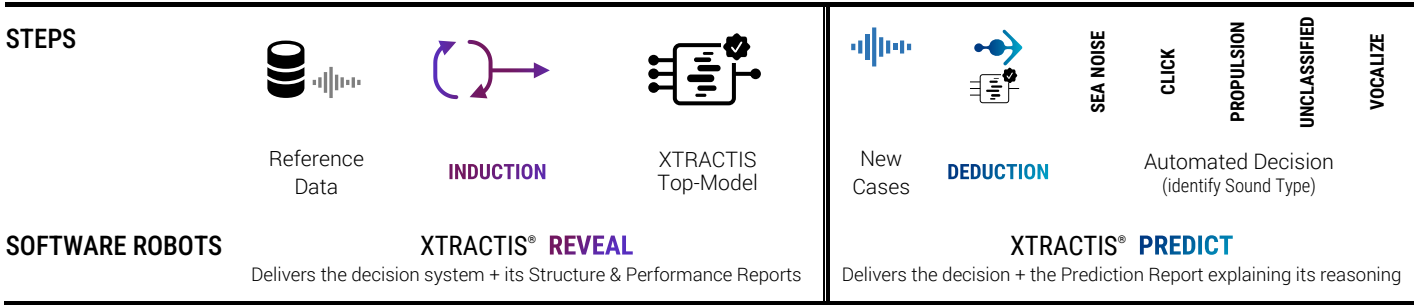
MODEL TYPE

Regression **Multinomial Classification** Binomial Classification Scoring

XTRACTIS-INDUCED DECISION SYSTEM

- Intelligible Model, Explainable Decisions**
 - ▶ The top-model is a decision system composed of 18 gradual rules without chaining.
 - ▶ Each rule uses from 1 to 4 predictors among the 9 variables that XTRACTIS automatically identified as significant (out of the 23 Potential Predictors).
 - ▶ Only a few rules are triggered at a time to compute the decision.
- High Predictive Capacity** It has an Excellent 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 2,000 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 4 possible operators into a **College of Virtual Experts (CVE)**.
- Among the 8,000 induced CVEs, the top-CVE with the best predictive performance remains complex: 867 rules share 14 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 13,920 new cases simulated by deduction from the top-CVE, around the 464 original learning cases but distinct from them.
- We apply 2,000 induction strategies to the same single partition of this new dataset (34% Training | 33% Validation | 33% Test): XTRACTIS induces 2,000 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

202,000 IVEs

Criterion for the induction optimization

Average F₁-Score

Validation criterion for the top-model selection

Average F₁-Score

Duration of the process @ Induction Speed FP64

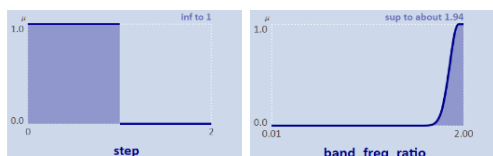
23 hours @ 1.13 Tflops

TOP-MODEL STRUCTURE

The top-IVE has an EXCELLENT intelligibility as it has **18 rules** combining **9 predictors**, with 2.6 predictors per rule on average. Its 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

- 9 continuous sound features (out of 23)
- Ranked by individual contribution (5 strong signals, 3 medium signals & 1 weak signal): #1 **max_ontime1** /.../ #9 **band_time_ratio**
- Labeled by fuzzy and binary classes
Examples: **binary interval** "inf to 1";
fuzzy interval "sup to about 1.94"



RULES

- 18 connective fuzzy rules without chaining (aggregated into 5 disjunctive fuzzy rules)
- 1 to 4 predictors per rule (on average, 2.6 predictors per rule)
- Example:
fuzzy rule R1 uses 3 predictors and concludes "Sea noise". 17 other fuzzy rules complete this model.

```

IF step IS inf to 1
AND band_freq_ratio IS sup to about 1.94
AND max_ontime1 IS about [1.1 ; 2.9]
THEN Sound Type IS Sea noise
    
```

Literally, the analyzed sound is a Sea noise if the step is 1 or 0 and the band/frequency ratio is above about 1.94 and the max_ontime1 is between 1.1 and 2.9.

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			464 Original Points
	34% Training (Synthetic Data)	33% Validation (Synthetic Data)	33% Test (Synthetic Data)	
Average F ₁ -Score	98.03%	97.79%	97.45%	97.65%
Classification Error	1.06%	1.22%	1.24%	1.08%

REAL
External Test
93.71%
3.53%

EXPLAINED PREDICTIONS FOR 3 UNKNOWN CASES

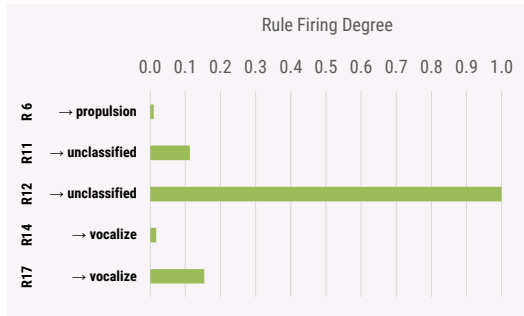
CASE

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

Sample #173	
step	1
iband	20
band_freq_ratio	0.03
band_time_ratio	631
nharmonic*	0 (1.89% OOR)
max_freq	6,761
max_ontime1	0.3
nreliance*	0 (2.38% OOR)
time_base2	0.3
Actual Value	unclassified

DEDUCTIVE INFERENCE OF RULES

For this sample, 5 rules are triggered:
R12 is fired at 1.000 and **R11** at 0.113 to conclude UNCLASSIFIED, and
R17 is fired at 0.154 and **R14** at 0.017 to conclude VOCALIZE, and
R6 is fired at 0.010 to conclude PROPULSION.
 The 13 other rules are not activated.



REAL-TIME DECISION

NUMBER OF TRIGGERED RULES

5 / 18

FUZZY PREDICTION

{ **unclassified** | 1.000,
vocalize | 0.154,
propulsion | 0.010 }

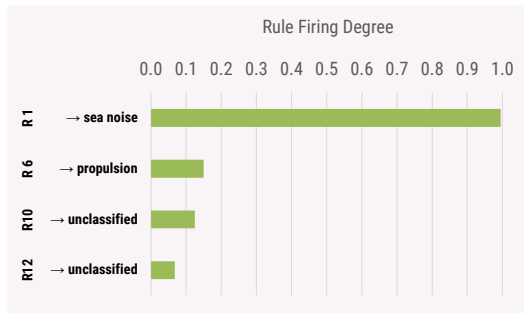
FINAL PREDICTION

{ **unclassified** }

The system delivers the correct decision compared to the actual case

Sample #498	
step	1
iband	21
band_freq_ratio*	2.00 (< 0.01% OOR)
band_time_ratio	86.9
nharmonic*	0 (1.89% OOR)
max_freq	129
max_ontime1	3
nreliance*	0 (2.38% OOR)
time_base2*	0.0 (< 0.01% OOR)
Actual Value	sea noise

For this sample, 4 rules are triggered:
R1 is fired at 0.995 to conclude SEA NOISE, and
R6 is fired at 0.150 to conclude PROPULSION, and
R10 is fired at 0.125 and **R12** at 0.068 to conclude UNCLASSIFIED.
 The 14 other rules are not activated.



NUMBER OF TRIGGERED RULES

4 / 18

FUZZY PREDICTION

{ sea noise | 0.995,
propulsion | 0.150,
unclassified | 0.125 }

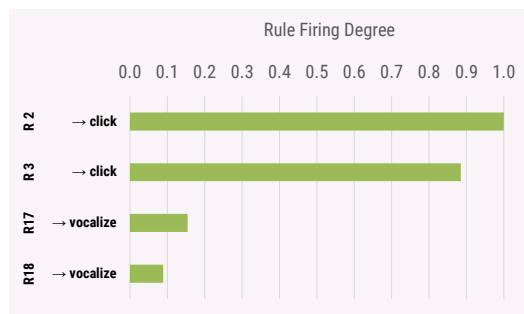
FINAL PREDICTION

{ **sea noise** }

The system delivers the correct decision compared to the actual case

Sample #739	
step	0
iband	1,177
band_freq_ratio	0.94
band_time_ratio	1190000
Nharmonic*	0 (1.89% OOR)
max_freq	14,772
max_ontime1	0
nreliance	6
time_base2	0.1
Actual Value	click

For this sample, 4 rules are triggered:
R2 is fired at 1.000 and **R3** at 0.885 to conclude CLICK, and
R17 is fired at 0.154 and **R18** at 0.089 to conclude VOCALIZE.
 The 14 other rules are not activated.



NUMBER OF TRIGGERED RULES

4 / 18

FUZZY PREDICTION

{ **click** | 1.000,
vocalize | 0.154 }

FINAL PREDICTION

{ **click** }

The system delivers the correct decision compared to the actual case

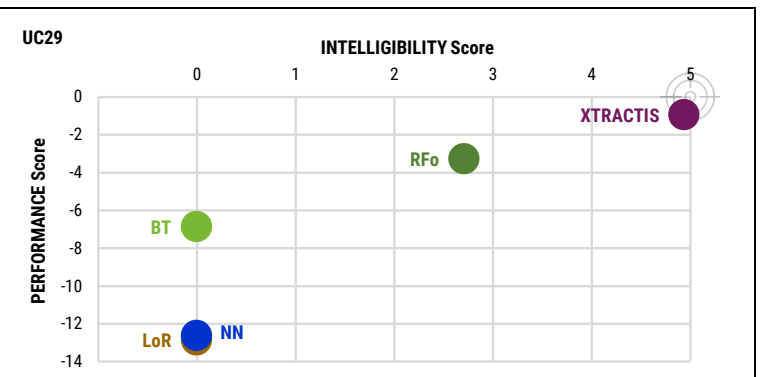
*Predictor value outside the variation range of the model but inside the allowed extrapolation range. XTRACTIS will refuse to give a result for an extrapolation far from the allowed extrapolation range.

TOP-MODELS BENCHMARK

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
MODELS RELEASE	2024/07	2024/09	2024/07	2024/07	2024/07
ALGORITHM VERSION	XTRACTIS REVEAL v. 13.0.51395	XTRACTIS BENCHMARK module embedding Python 3.9.10 Scikit-Learn 1.3.0 LightGBM 3.3.2 TensorFlow 2.10.0 Keras 2.10.0			
CROSS-VALIDATION TECHNIQUE	20 × 5 folds for each CVE model. Then 1-Split Validation for each IVE model: 34% Training 33 % Validation 33% Test	20 × 5 folds for each CVE model			
NUMBER OF EXPLORED STRATEGIES⁽¹⁾	2,000 induction strategies for the CVE on Training / Validation data. 4 aggregation operators tested. 2,000 induction strategies for the IVE on synthetic data	2,000 ML strategies on Training / Validation data. Aggregation operator: Relative Majority			
TOP-MODEL SELECTION⁽²⁾	Top-CVE with Simple Majority aggregator, selected among the 8,000 CVEs. Then Top-IVE among 2,000 IVEs	Top-CVE selected among 2,000 CVEs then single model obtained by applying best CVE strategy on 100% of the Learning Dataset			

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
NUMBER OF PREDICTORS (out of 23 Potential Predictors)	9	26 1 nominal predictor with 4 modalities split into 4 predictors	21	22	26 1 nominal predictor with 4 modalities split into 4 predictors
AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	2.6 per rule	26.0 per equation	3.4 per rule	3.1 per rule	30.5 per equation
STRUCTURE OF THE DECISION SYSTEM	18 fuzzy rules without chaining (aggregated into 5 disjunctive fuzzy rules) Only a few rules are triggered at a time to compute a decision	5 linear equations	45 trees without chaining 365 binary rules	5 chains of 28 trees each 976 binary rules Tree #N corrects the error of the N-1 previous trees	5 hidden layers 133 hidden nodes 138 equations 133 unintelligible synthetic variables, in addition to the 26 original predictors

	Random ⁽³⁾	XTRACTIS	LoR	RFo	BT	NN
INTELLIGIBILITY Score⁽⁴⁾		4.94	0.00	2.71	0.00	0.00
CVE Real Perf. (Average F ₁ -Score) in External Test	57.65	94.93	82.40	88.40	88.40	85.36
Gap to CVE Leader in External Test	-37.28	0.00	-12.53	-6.53	-6.53	-9.57
IVE Real Perf. (Average F ₁ -Score) in External Test	57.65	93.71	82.40	95.60	88.40	79.94
Gap to IVE Leader in External Test	-37.95	-1.89	-13.20	0.00	-7.20	-15.66
Average Real Performance in External Test	57.65	94.32	82.40	92.00	88.40	82.65
PERFORMANCE Score⁽⁴⁾		-0.95	-12.87	-3.27	-6.86	-12.62



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

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

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 trees per chain, here for BT only):

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

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

- 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	Average Sensitivity	Min. PPV	Average PPV	Min. F ₁ -Score	Average F₁-Score	Weighted Av. F ₁ -Score	Refusal
RANDOM MODEL									
<i>Number of Random Permutations (P-value) = 100,000 (0.001)</i>									
<i>Performance against chance</i>									
	42.35%	23.08%	36.50%	23.08%	36.50%	23.08%	57.65%	62.96%	
XTRACTIS TOP-MODEL									
CVE - Descriptive Performance (Training)	0.22%	99.67%	99.93%	96.67%	99.33%	98.31%	99.63%	99.79%	0 (0.00%)
CVE - Predictive Performance (Validation)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0 (0.00%)
CVE - Real Performance (External Test)	3.53%	83.33%	95.93%	83.33%	94.07%	83.33%	94.93%	96.49%	0 (0.00%)
IVE - Descriptive Performance (Training)	1.06%	96.79%	98.72%	93.26%	97.36%	94.99%	98.03%	98.95%	0 (0.00%)
IVE - Predictive Performance (Validation)	1.22%	94.93%	98.15%	94.64%	97.43%	94.78%	97.79%	98.78%	0 (0.00%)
IVE - Real Performance (Test)	1.24%	95.22%	97.60%	94.36%	97.31%	96.08%	97.45%	98.76%	0 (0.00%)
IVE - Real Performance (464 original points)	1.08%	93.10%	97.67%	90.00%	97.80%	94.74%	97.65%	98.92%	0 (0.00%)
IVE - Real Performance (External Test)	3.53%	66.67%	92.96%	88.89%	95.98%	80.00%	93.71%	96.29%	0 (0.00%)
LOGISTIC REGRESSION TOP-MODEL									
CVE - Descriptive Performance (Training)	9.48%	83.33%	92.36%	50.91%	85.40%	66.67%	87.25%	91.26%	
CVE - Predictive Performance (Validation)	10.78%	77.78%	88.67%	48.08%	81.88%	61.73%	84.04%	89.97%	
CVE - Real Performance (External Test)	15.29%	75.00%	87.59%	55.56%	80.67%	66.67%	82.40%	85.12%	
IVE - Descriptive Performance (Training)	9.27%	83.33%	92.04%	53.85%	84.43%	69.14%	86.87%	91.34%	
IVE - Real Performance (External Test)	15.29%	75.00%	87.59%	55.56%	80.67%	66.67%	82.40%	85.12%	
RANDOM FOREST TOP-MODEL									
CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Predictive Performance (Validation)	2.16%	83.33%	94.20%	91.49%	97.26%	90.91%	95.56%	97.83%	
CVE - Real Performance (External Test)	5.88%	50.00%	84.63%	88.89%	96.37%	66.67%	88.40%	93.56%	
IVE - Descriptive Performance (Training)	0.22%	99.67%	99.93%	96.67%	99.33%	98.31%	99.63%	99.79%	
IVE - Real Performance (External Test)	2.35%	75.00%	94.63%	88.89%	97.41%	85.71%	95.60%	97.60%	
BOOSTED TREE TOP-MODEL									
CVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Predictive Performance (Validation)	1.51%	88.89%	95.20%	96.30%	98.49%	92.86%	96.75%	98.47%	
CVE - Real Performance (External Test)	5.88%	50.00%	84.63%	88.89%	96.37%	66.67%	88.40%	93.56%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	5.88%	50.00%	84.63%	88.89%	96.37%	66.67%	88.40%	93.56%	
NEURAL NETWORK TOP-MODEL									
CVE - Descriptive Performance (Training)	0.65%	94.44%	98.13%	97.83%	99.21%	97.14%	98.65%	99.35%	
CVE - Predictive Performance (Validation)	1.08%	96.55%	98.67%	96.55%	98.61%	96.55%	98.63%	98.92%	
CVE - Real Performance (External Test)	8.24%	50.00%	81.05%	88.89%	94.48%	66.67%	85.36%	91.28%	
IVE - Descriptive Performance (Training)	1.51%	88.89%	97.07%	89.80%	97.01%	93.62%	96.93%	98.51%	
IVE - Real Performance (External Test)	12.94%	50.00%	80.07%	60.00%	82.80%	54.55%	79.94%	87.26%	

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 (2024-2025). XTRACTIS® the General Reasoning AI for Trusted Decisions. Use Case #29 | Naval Security: Identification of Underwater Sounds (Virtual Golden Ear) – Benchmark vs. Logistic Regression, Random Forest, Boosted Tree & Neural Network. IDAGRAI LABS, June 2025, v2.0, Compiegne, France, 6p.