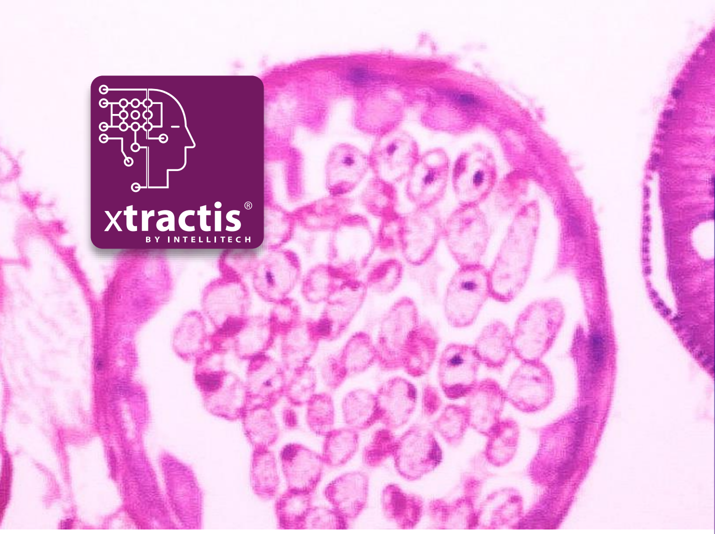


Precision Medicine

ANATOMOPATHOLOGICAL DIAGNOSIS OF BREAST CANCER

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

UC#04 — 2024/03 (v3.1) xtractis.ai

PROBLEM DEFINITION

GOAL

Design an AI-based decision system that accurately and instantly makes a rational medical diagnosis of breast cancer from microscopic images of patient mammary cells.

PROS & BENEFITS

- Identify the cellular characteristics involved in cancer and enhance medical knowledge by helping pathologists and oncologists understand the causal relationships between specific cell features, 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:
Dr. William H. Wolberg,
W. Nick Street,
Olvi L. Mangasarian –
University of Wisconsin
[UCI ML Repository]

Variable to Predict: The model makes the diagnosis of breast cells as **BENIGN** | **MALIGNANT**.

Potential Predictors: 30 variables are topological and geometric attributes of mammary cells:
Radius, Texture, Perimeter, Area, Concavity, Symmetry, Smoothness....

Observations: 569 images of mammary cells from patients with or without cancer.

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: 483 | 84.89% cases 80% for Training, 20% for Validation	
BENIGN	MALIGNANT
303 | 62.73%	180 | 37.27%

External Test Dataset: 86 | 15.11% cases	
BENIGN	MALIGNANT
54 | 62.79%	32 | 37.21%

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 7 gradual rules without chaining aggregated into 2 disjunctive rules.
- Each rule uses from 2 to 5 predictors among the 13 variables that XTRACTIS automatically identified as significant (out of the 30 attributes of mammary cells describing each image).
- 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).

☑ **Efficient AI System** It computes real-time predictions up to 70,000 decisions/second, offline or online (API).

XTRACTIS PROCESS

STEPS	Reference Data	**INDUCTION**	XTRACTIS Top-Model	New Cases	**DEDUCTION**	BENIGN / MALIGNANT Automated Decision (detect cancer)
SOFTWARE ROBOTS	XTRACTIS® **REVEAL** Delivers the decision system + its Structure & Performance Reports			XTRACTIS® **PREDICT** Delivers the decision + the Prediction Report explaining its reasoning		

TOP-MODEL INDUCTION

INDUCTION PARAMETERS

Powered by:
XTRACTIS® REVEAL
v12.1.42004

1. We launch 2,000 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.
2. 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).
3. Among the 6,000 induced CVEs, the top-CVE with the best predictive performance remains complex: 1,338 rules sharing 30 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:

4. We build a synthetic dataset composed of 24,150 new cases simulated by deduction from the top-CVE, around the 483 original learning cases but distinct from them.
5. We apply 2,000 induction strategies to the same single 34% Training | 33% Validation | 33% Test partition of this new dataset: XTRACTIS induces 2,000 IVEs.
6. The top-IVE selected is the one that is the most intelligible while being as efficient as the top-CVE.

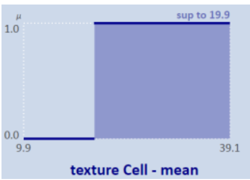
Total number of induced unitary models	Criterion for the induction optimization	Validation criterion for the top-model selection	Duration of the process (Induction Power FP64)
402,000 IVEs	F_1-Score	F_1-Score	27 hours (1 Tflops)

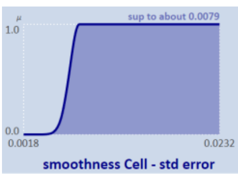
TOP-MODEL STRUCTURE

The top-IVE model has an excellent intelligibility as it combines the 13 predictors into 7 rules with 3 predictors 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

- 13 cell characteristics out of 30
- Ranked by individual contribution (4 medium & 9 weak signals): #1 Perimeter Cell…/ #2 Concave points Cell… / …
- Labeled by fuzzy and binary classes
  Examples: **binary interval** "sup. to 19.9"
  fuzzy interval "sup. to about 0.0079"





RULES

- 7 connective fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules)
- 2 to 5 predictors per rule (on average, 3.1 predictors per rule)
- Example: fuzzy rule **R5** uses 3 predictors and concludes **MALIGNANT**. 6 other rules complete this model, including 2 binary rules.

IF	texture Cell - mean	IS	sup. to 19.9
AND	smoothness Cell - std error	IS	sup. to ~0.0079
AND	area Cell - mean_3_largest	IS	sup. to 797
THEN	Diagnosis	IS	MALIGNANT

Literally, the cells image indicates a malignant tumor diagnosis if the "texture Cell – mean" is superior to 19.9, and the "smoothness Cell - std error" is above around 0.0079, and the "area Cell - mean_3_largest" is superior to 797.

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.

	Synthetic Data			Reference Data
Performance	DESCRIPTIVE	PREDICTIVE	REAL	REAL
Dataset	34% Training	33% Validation	33% Test	External Test
F_1-Score	99.19%	99.14%	98.95%	98.46%
Classification Error	0.13%	0.21%	0.13%	1.16%

EXPLAINED PREDICTIONS FOR 2 UNKNOWN CASES

Powered by: XTRACTIS® PREDICT v12.1.42004

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

DEDUCTIVE INFERENCE OF RULES

AUTOMATED DECISION

PATIENT #881094802

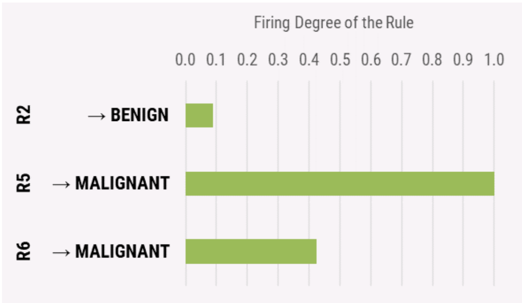
actual value = MALIGNANT	
texture Cell - mean	25.6
smoothness Cell - mean	0.101
concavity Cell - mean	0.168
concave points Cell - mean	0.066
symmetry Cell - mean	0.131
smoothness Cell - std error*	0.0311
radius Cell - mean_3_largest	18.1
texture Cell - mean_3_largest	28.1
perimeter Cell - mean_3 largest	120
area Cell - mean_3_largest	1,021
smoothness Cell - mean_3_largest	0.124
concavity Cell - mean_3_largest	0.28
concave points Cell - mean_3_largest	0.110

Real Time

For this patient, 3 rules are triggered:

R5 is fired at 1.000, **R6** at 0.424, and **R2** at 0.089.

R1, R3, R4, and R7 are not activated.



NUMBER OF TRIGGERED RULES
3 / 7

FUZZY PREDICTION
{ MALIGNANT | 1.000,
BENIGN | 0.089 }

FINAL PREDICTION
{ MALIGNANT }

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

MALIGNANT

PATIENT #866458

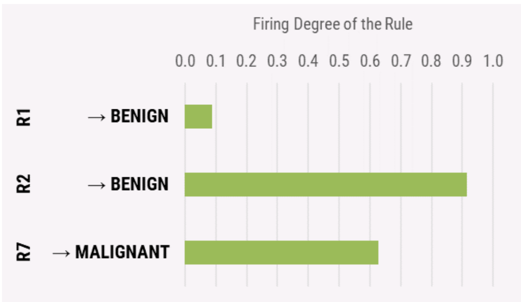
actual value = BENIGN	
texture Cell - mean	16.4
smoothness Cell - mean	0.115
concavity Cell - mean	0.114
concave points Cell - mean	0.085
symmetry Cell - mean	0.200
smoothness Cell - std error	0.0090
radius Cell - mean_3_largest	16.1
texture Cell - mean_3_largest	18.3
perimeter Cell - mean_3 largest	106
area Cell - mean_3_largest	763
smoothness Cell - mean_3_largest	0.139
concavity Cell - mean_3_largest	0.20
concave points Cell - mean_3_largest	0.142

Real Time

For this patient, 3 rules are triggered:

R2 is fired at 0.914, **R7** at 0.629, and **R1** at 0.088.

R3, R4, R5, and R6 are not activated.



NUMBER OF TRIGGERED RULES
3 / 7

FUZZY PREDICTION
{ BENIGN | 0.914,
MALIGNANT | 0.629 }

FINAL PREDICTION
{ BENIGN }

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

BENIGN

**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: DECISION STRUCTURE & INTELLIGIBILITY × PERFORMANCE SCORES

MODELING PARAMETERS

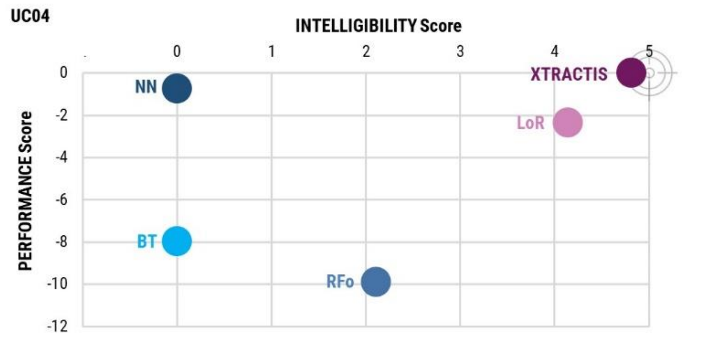
	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
MODELS RELEASE	2022/02	2022/09	2022/08	2022/07	2022/09
ALGORITHM VERSION	XTRACTIS **REVEAL** 12.1.42004	Python 3.7; Scikit-learn 1.0.2	Python 3.6; LightGBM 2.2.2	Python 3.6; LightGBM 2.2.2	Python 3.6; TensorFlow 2.6.2, Keras 2.6.0
CROSS-VALIDATION TECHNIQUE	40×5 folds for each CVE model. Then 1-Split Validation for each IVE model: 34% Training | 33% Validation | 33% Test	40×5 folds for each CVE model	40×5 folds for each CVE model	40×5 folds for each CVE model	40×5 folds for each CVE model
NUMBER OF EXPLORED STRATEGIES(1)	2,000 induction strategies for the CVE on Training / Validation data. 2,000 induction strategies for the IVE on synthetic data	2,000 data analysis strategies on Training / Validation data	1,020 ML strategies on Training / Validation data	1,000 ML strategies on Training / Validation data	2,000 ML strategies on Training / Validation data
TOP-MODEL SELECTION(2)	Top-CVE among 6,000 CVEs. Then Top-IVE among 2,000 IVEs	Top-CVE among 2,000 CVEs	Top-CVE among 1,020 CVEs	Top-CVE among 1,000 CVEs	Top-CVE among 2,000 CVEs
		then single model obtained by applying best CVE strategy on 100% of the Learning Dataset			

TOP-MODEL STRUCTURE

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
NUMBER OF PREDICTORS (out of 30 Potential Predictors)	**13**	**5**	**28**	**30**	**30**
AVERAGE NUMBER OF PREDICTORS PER RULE / EQUATION	**3.1** per rule	**5** per equation	**4.3** per rule	**3.6** per rule	**34.2** per equation
STRUCTURE OF THE DECISION SYSTEM	**7** fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules) Only a few rules are triggered at a time to compute a decision	**1** linear equation	**22** trees without chaining **379** binary rules	**1** chain of **34** trees **359** binary rules Tree #N corrects the error of the N-1 previous trees	**2** hidden layers | **64** hidden nodes **65** equations 64 unintelligible synthetic variables

TOP-MODEL SCORES

	Random(3)	XTRACTIS	LoR	RFo	BT	NN
INTELLIGIBILITY Score(4)		**4.82**	**4.14**	**2.11**	**0.00**	**0.00**
CVE Real Performance (F_1-Score) in External Test		96.99	95.38	91.43	90.91	95.52
Gap to CVE Leader in External Test		**0.00**	**-1.61**	**-5.56**	**-6.08**	**-1.47**
IVE Real Performance (F_1-Score) in External Test	*37.81*	98.46	95.38	84.21	88.57	98.46
Gap to IVE Leader in External Test		**0.00**	**-3.08**	**-14.25**	**-9.89**	**0.00**
Average Real Performance in External Test	*37.81*	97.73	95.38	87.82	89.74	96.99
PERFORMANCE Score(4)		**0.00**	**-2.35**	**-9.90**	**-7.99**	**-0.73**



(1) For all algos: on the same Learning Dataset. All Models are optimized according to their Validation F_1-Score.
(2) All top-models are selected according to their Validation F_1-Score while checking that it remains close to their Training F_1-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	$i \in [1 ; n]$ n = number of AI Techniques benchmarked in terms of data-driven modeling = 5
Benchmark #k	B_k	$k \in [1 ; p]$ p = number of Benchmarks for the Use Case $\in \{1, 2, 3\}$

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; F_1-Score for a Binomial Classification; Average F_1-Score or Average F_2-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: $PS(T_i, B_k) = PC(T_i, B_k) - Best\_PC(B_k)$.

Performance Score of T_i

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

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):
  $$Pen1(T_i) = \min(0\ , 1 - \log_{10} number\ of\ predictors)$$
  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):
  $$Pen2(T_i) = \min\left(0\ , 0.01 - \frac{average\ number\ of\ rules\ or\ equations\ per\ modality\ to\ predict}{100}\right)$$
  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):
  $$Pen3(T_i) = \min\left(0\ , \frac{9 - 3 \times average\ number\ of\ predictors\ per\ rule\ or\ equation}{7}\right)$$
  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):
  $$Pen4(T_i) = \min(0\ , 1 - number\ of\ chained\ trees)$$
  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):
  $$Pen5(T_i) = -5$$

Intelligibility Score of T_i

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

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 Ti 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								
Number of Random Permutations (P-value) = 100,000 (0.001%)								
Performance against chance (External Test)	*45.43%*	*37.81%*					*37.81%*	
XTRACTIS TOP-MODEL								
CVE - Descriptive Performance (Training)	0.83%	97.78%	97.78%	100.00%	100.00%	98.70%	98.88%	0 (0.00%)
CVE - Predictive Performance (Validation)	0.62%	98.33%	98.33%	100.00%	100.00%	99.02%	99.16%	0 (0.00%)
CVE - Real Performance (External Test)	2.33%	96.30%	100.00%	96.30%	94.12%	100.00%	96.99%	0 (0.00%)
IVE - Descriptive Performance (Training)	0.13%	99.85%	99.90%	99.85%	99.73%	99.94%	99.19%	0 (0.00%)
IVE - Predictive Performance (Validation)	0.21%	99.78%	99.79%	99.78%	99.62%	99.88%	99.14%	0 (0.00%)
IVE - Real Performance (Test)	0.13%	98.84%	99.93%	99.84%	99.73%	99.96%	98.95%	0 (0.00%)
IVE - Real Performance (483 original points)	0.83%	97.78%	97.78%	100.00%	100.00%	98.70%	98.88%	0 (0.00%)
IVE - Real Performance (External Test)	1.16%	98.15%	100.00%	98.15%	96.97%	100.00%	98.46%	0 (0.00%)
LOGISTIC REGRESSION TOP-MODEL								
CVE - Descriptive Performance (Training)	3.11%	95.00%	95.00%	98.02%	96.61%	97.06%	95.80%	
CVE - Predictive Performance (Validation)	1.66%	97.22%	97.22%	99.01%	98.31%	98.36%	97.77%	
CVE - Real Performance (External Test)	3.49%	96.30%	96.88%	96.30%	93.94%	98.11%	95.38%	
IVE - Descriptive Performance (Training)	2.69%	95.56%	95.56%	98.35%	97.18%	97.39%	96.36%	
IVE - Real Performance (External Test)	3.49%	96.30%	96.88%	96.30%	93.94%	98.11%	95.38%	
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)	1.86%	96.67%	96.67%	99.01%	98.31%	98.04%	97.48%	
CVE - Real Performance (External Test)	6.98%	88.88%	100.00%	88.88%	84.21%	100.00%	91.43%	
IVE - Descriptive Performance (Training)	0.21%	99.67%	100.00%	99.67%	99.45%	100.00%	99.72%	
IVE - Real Performance (External Test)	13.95%	77.77%	100.00%	77.77%	72.73%	100.00%	84.21%	
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)	1.24%	96.67%	96.67%	100.00%	100.00%	98.06%	98.31%	
CVE - Real Performance (External Test)	6.97%	92.59%	93.75%	92.59%	88.24%	96.15%	90.91%	
IVE - Descriptive Performance (Training)	0.21%	99.44%	99.44%	100.00%	100.00%	99.67%	99.72%	
IVE - Real Performance (External Test)	9.30%	87.04%	96.87%	87.04%	81.58%	97.92%	88.57%	
NEURAL NETWORK TOP-MODEL								
CVE - Descriptive Performance (Training)	2.69%	97.22%	97.22%	97.36%	95.63%	98.33%	96.42%	
CVE - Predictive Performance (Validation)	2.28%	97.22%	97.22%	98.02%	96.69%	98.34%	96.95%	
CVE - Real Performance (External Test)	3.49%	94.44%	100.00%	94.44%	91.43%	100.00%	95.52%	
IVE - Descriptive Performance (Training)	1.45%	97.22%	97.22%	99.34%	100.00%	100.00%	98.04%	
IVE - Real Performance (External Test)	1.16%	98.15%	100.00%	98.15%	96.97%	100.00%	98.46%	