Cyber Security

LOG-BASED DETECTION OF CYBER INTRUSIONS (DARPA)

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

UC#06 — 2024/06 (v5.2)

xtractis.ai

PROBLEM DEFINITION

GOAL

Design an AI-based decision system that accurately diagnoses an intrusion on a computer network from features of the connection logs, to instantly execute the appropriate rational action.

PROS & BENEFITS

- Identify the characteristics of logs defining a cyber intrusion. Enhance expert knowledge by helping cybersecurity specialists understand the causal relationships between specific log features, their combination, and the existence of an intrusion.
- Help IT detect cyberattacks as early as possible and understand the underlying strategy of the attacker in order to consider measures to thwart future attacks.
- Avoid many false alarms thanks to transparent diagnosis, in a context of increasing number of attacks with the use of open-source AI algorithms.

REFERENCE DATA

Source:
Cyber Systems and Technology group of MIT Lincoln Laboratory, DARPA ITO, Air Force Research Laboratory [UCI Machine Learning Repository].

Variable to Predict

The model predicts the connection state: **NORMAL** | **INTRUSION**.

Potential Predictors

41 Variables characterizing each log:
duration, protocol type, network service, number of data bytes from source to destination, flag status of connection...

Observations

1,074,983 connection logs on the US Air Force military computer network. Each log is associated with a normal activity or an attack. Data are divided into

- a Learning Dataset for model induction using Training, Validation and Test Datasets,
- and an External Test Dataset (ETD#1) with an environment close to the learning one to check the top-model's performance on real data and for benchmarking.

An additional dataset of 70,874 connections corresponding to a network environment that has strongly changed is used as a second External Test Dataset (ETD#2).

All duplicates were removed from the reference dataset to avoid biasing performance assessment.

Learning Dataset: 859,984 logs | 80% 70% for Training, 15% for Validation, 15% for Test	
NORMAL	INTRUSION
650,239 | 75.61%	209,750 | 24.39%

ETD#1: 214,999 logs | 20%	
NORMAL	INTRUSION
162,559 | 75.61%	52,438 | 24.39%

ETD#2: 70 874 logs	
NORMAL	INTRUSION
47,578 | 67.13%	23,296 | 32.87%

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 25 gradual rules without chaining.
- Each rule uses from 3 to 8 predictors among the 26 log characteristics that XTRACTIS automatically identified as significant (out of the 41 Potential Predictors).
- Only a few rules are triggered at a time to compute the decision.

☑ **High Predictive Capacity**

It has a very good to 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

STEPS	Logs Reference Data	**INDUCTION**	XTRACTIS Top-Model	Logs New Cases	**DEDUCTION**	NORMAL / INTRUSION Automated Decision (detect attack)	
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
v12.1.42925

1. We launch 500 inductive reasoning strategies; each strategy is applied to the same single partition of the learning dataset (70% Training / 15% Validation / 15% Test) to get a reliable assessment of the descriptive and predictive performances, respectively from Training and Validation Datasets.
2. Each strategy thus generates one unitary model called **Individual Virtual Expert** (IVE).
3. Among the 500 induced models, the top-IVE selected is the one that has the best predictive performance, close to its descriptive performance, and with the best intelligibility, i.e., with the fewer predictors and rules.

Total number of induced unitary models
500 IVEs

Criterion for the induction optimization
F_1-Score

Validation criterion for the top-model selection
F_1-Score

Duration of the process @ Induction Speed FP64
4 days @ 24 Tflops

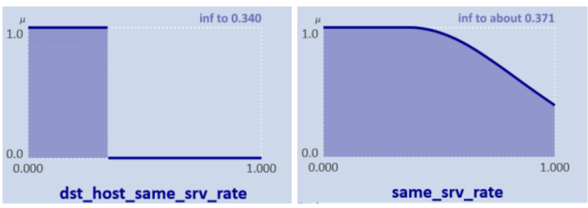
TOP-MODEL STRUCTURE

The top-IVE has a very good intelligibility as it has **25 rules** combining **26 predictors**, with 6.7 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

- 26 log characteristics (out of 41)
- 23 continuous + 3 nominal variables
- Ranked by impact significance
  (4 strong, 11 medium & 11 weak signals):
  #1 src_bytes_1450Clip … / #2 duration_3Clip …
- Labeled by fuzzy and binary classes
  Examples: **binary interval** "inf to 0.340";
  fuzzy interval "inf to about 0.371"



RULES

- 25 connective fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules)
- 3 to 8 predictors per rule (on average, 5.6 predictors per rule)
- Example: **fuzzy rule R21** uses 3 predictors to conclude "INTRUSION". 24 other fuzzy rules complete this model.

IF	same_srv_rate	IS	inf to about 0.371
AND	dst_host_same_srv_rate	IS	inf to 0.340
AND	src_bytes_1450Clip	IS	{0}
THEN	Connection	IS	INTRUSION

Literally, the connection is an intrusion if the rate of connections to the same service of the same target during the last 2 seconds is inferior to around 37% and the rate of connections, among the last 100, to the same service of the same target is inferior to 34%, and the number of data bytes sent by the source to the target is zero.

TOP-MODEL PERFORMANCE

The top-IVE performances, measured in Training/Validation/Test, then in External Test on ETD#1 and ETD#2, guarantee the model's predictive and real performances.

Performance Type	DESCRIPTIVE	PREDICTIVE	REAL	REAL	REAL
Dataset	70% Training	15% Validation	15% Test	ETD #1	ETD #2
F_1-Score	**99.93%**	**99.94%**	**99.91%**	**99.93%**	**92.05%**
Classification Error	0.03%	0.03%	0.05%	0.04%	4.93%

EXPLAINED PREDICTIONS FOR 3 UNKNOWN CASES

Powered by: XTRACTIS® PREDICT v12.1.42925

NEW CASE (from the External Test Dataset, i.e., not included in the Learning Dataset) — **DEDUCTIVE INFERENCE OF RULES** — **AUTOMATED DECISION**

LOG V_161144

actual value = INTRUSION

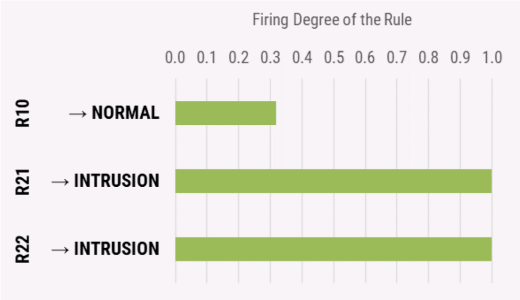
rerror_rate	1.000
same_srv_rate	0.030
diff_srv_rate	0.060
dst_host_count	255
dst_host_srv_count	9
dst_host_same_srv_rate	0.040
dst_host_diff_srv_rate	0.060
dst_host_same_src_port_rate	0.000
…	…
duration_3Clip	0.00
src_bytes_1450Clip	0
srv_count_35Clip	9.0
protocol_typ	tcp
service	smtp
flag	RSTO

Real Time

For this connection, 3 rules are triggered:

R21 and **R22** at 1.000, **R10** at 0.381.

The 22 other rules are not activated.



NUMBER OF TRIGGERED RULES
3 / 25

FUZZY PREDICTION
{ INTRUSION | 1.000,
NORMAL | 0.381 }

FINAL PREDICTION
{ INTRUSION }

The system delivers the correct diagnosis compared to that given by the cyber expert:

INTRUSION

LOG V_100052

actual value = NORMAL

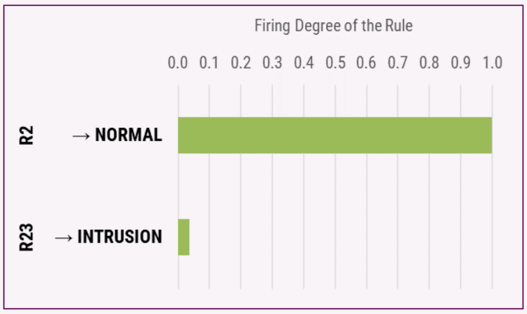
rerror_rate	0.000
same_srv_rate	1.000
diff_srv_rate	0.000
dst_host_count	28
dst_host_srv_count	11
dst_host_same_srv_rate	0.390
dst_host_diff_srv_rate	0.110
dst_host_same_src_port_rate	0.040
…	…
duration_3Clip	3.00
src_bytes_1450Clip	241
srv_count_35Clip	1.0
protocol_typ	tcp
service	ftp
flag	SF

Real Time

For this connection, 2 rules are triggered:

R2 at 1.000 and **R23** at 0.037.

The 23 other rules are not activated.



NUMBER OF TRIGGERED RULES
2 / 25

FUZZY PREDICTION
{ NORMAL | 1.000,
INTRUSION | 0.037 }

FINAL PREDICTION
{ NORMAL }

The system delivers the correct diagnosis compared to that given by the cyber expert:

NORMAL

LOG V_41490

actual value = NORMAL

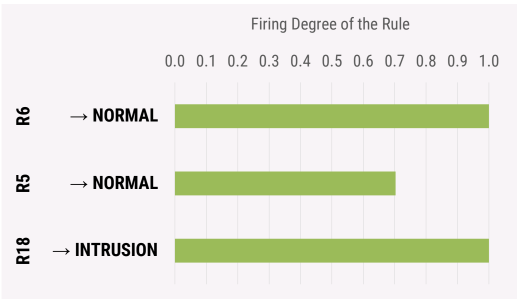
rerror_rate	0.000
same_srv_rate	1.000
diff_srv_rate	0.000
dst_host_count	12
dst_host_srv_count	12
dst_host_same_srv_rate	1.000
dst_host_diff_srv_rate	0.000
dst_host_same_src_port_rate	1.000
…	…
duration_3Clip	0.00
src_bytes_1450Clip	30
srv_count_35Clip	1.0
protocol_typ	icmp
service	ecr_i
flag	SF

Real Time

For this connection, 3 rules are triggered:

R6 and **R18** at 1.000, **R5** at 0.703.

The 22 other rules are not activated.



NUMBER OF TRIGGERED RULES
3 / 25

FUZZY PREDICTION
{ NORMAL | 1.000,
INTRUSION | 1.000 }

FINAL PREDICTION
REFUSAL

The system cannot deliver a valid diagnosis, so it refuses to decide.

This conflicting situation is a warning for cyber experts to analyze this log in depth.

More training data with situations near this log profile should strengthen the model in this decision space area.

TOP-MODELS BENCHMARK

MODELING PARAMETERS

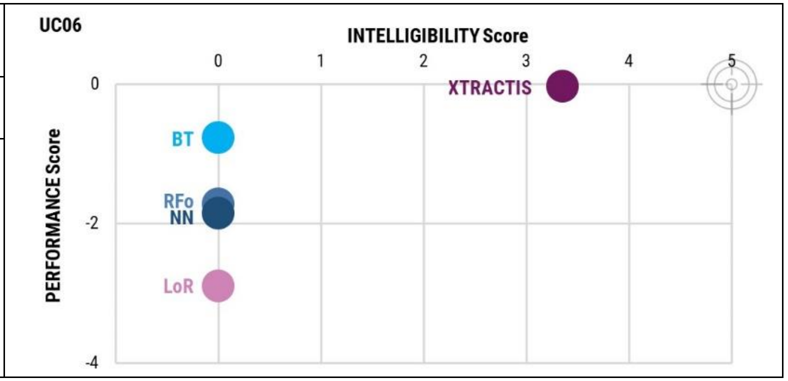
	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
MODELS RELEASE	2022/07	2022/09	2022/07	2022/07	2022/07
ALGORITHM VERSION	XTRACTIS **REVEAL** 12.1.42925	Python 3.7 | Scikit-Learn 1.0.2	Python 3.7 | LightGBM 2.2.2	Python 3.7 | LightGBM 2.2.2	Python 3.7 | TensorFlow 2.6.2 | Keras 2.6.0
CROSS-VALIDATION TECHNIQUE	All explored strategies for all algorithms use the same single-split of the Learning Dataset: 70% Training | 15% Validation | 15% Test				
NUMBER OF EXPLORED STRATEGIES(1)	500 induction strategies	500 data analysis strategies	500 ML strategies	500 ML strategies	500 ML strategies
TOP-MODEL SELECTION(2)	Top-IVE among 500 IVEs	Top-IVE among 500 IVEs	Top-IVE among 500 IVEs	Top-IVE among 500 IVEs	Top-IVE among 500 IVEs

TOP-MODEL STRUCTURE

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
NUMBER OF PREDICTORS (out of 41 Potential Predictors)	**26**	**32**	**36**	**32**	**122** 3 nominal variables are decomposed into 84 binary variables
AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	**5.6** per rule	**32.0** per equation	**9.0** per rule	**6.9** per rule	**68.5** per equation
STRUCTURE OF THE DECISION SYSTEM	**25** 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	**24** trees without chaining **3,023** binary rules	1 chain of **148** trees **8,393** binary rules Tree #N corrects the error of the N-1 previous trees	**4** hidden layers | **72** hidden nodes **73** equations 72 unintelligible synthetic variables

TOP-MODEL SCORES

	Random(3)	XTRACTIS	LoR	RFo	BT	NN
INTELLIGIBILITY Score(4)		**3.36**	**0.00**	**0.00**	**0.00**	**0.00**
IVE Real Performance (F_1-Score) in Test		99.91	98.95	99.89	99.98	99.89
Gap to Leader in Test		**-0.07**	**-1.03**	**-0.09**	**0.00**	**-0.09**
IVE Real Perf. (F_1-Score) in External Test #1	*24.90*	99.93	98.95	99.91	99.96	99.90
Gap to Leader in External Test #1		**-0.03**	**-1.01**	**-0.05**	**0.00**	**-0.06**
IVE Real Perf. (F_1-Score) in External Test #2	*33.65*	92.05	85.41	87.04	89.73	86.64
Gap to Leader in External Test #2		**0.00**	**-6.64**	**-5.01**	**-2.32**	**-5.41**
Average Real Performance	*29.28*	97.30	94.44	95.61	96.55	95.48
PERFORMANCE Score(4)		**-0.03**	**-2.89**	**-1.72**	**-0.77**	**-1.85**



(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 ∈ {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

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

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 trees per chain, here for BT only):
  $$Pen4(T_i) = \min(0\,, 1 - number\ of\ trees\ per\ chain)$$
  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):
  $$Pen5(T_i) = -5$$

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 in External Test 1	*36.63%*	*24.90%*					***24.90%***	
Performance against chance in External Test 2	*43.62%*	*33.65%*					***33.65%***	
XTRACTIS TOP-MODEL								
Descriptive Performance (Training)	0.03%	99.92%	99.92%	99.98%	99.95%	99.97%	99.93%	1 408 (0.23%)
Predictive Performance (Validation)	0.03%	99.92%	99.92%	99.99%	99.96%	99.98%	99.94%	297 (0.23%)
Real Performance (Test)	0.05%	99.89%	99.89%	99.98%	99.92%	99.96%	99.91%	303 (0.23%)
Real Performance (External Test 1)	0.04%	99.92%	99.92%	99.98%	99.93%	99.97%	**99.93%**	501 (0.23%)
Real Performance (External Test 2)	4.93%	86.44%	86.44%	99.32%	98.43%	93.70%	**92.05%**	803 (1.13%)
LOGISTIC REGRESSION TOP-MODEL								
Descriptive Performance (Training)	0.53%	98.60%	98.60%	99.75%	99.21%	99.55%	98.90%	
Predictive Performance (Validation)	0.52%	98.60%	98.60%	99.76%	99.26%	99.55%	98.93%	
Real Performance (Test)	0.51%	98.64%	98.64%	99.76%	99.26%	99.56%	98.95%	
Real Performance (External Test 1)	0.51%	98.65%	98.65%	99.76%	99.25%	99.57%	**98.95%**	
Real Performance (External Test 2)	8.45%	75.28%	75.28%	99.51%	98.69%	89.15%	**85.41%**	
RANDOM FOREST TOP-MODEL								
Descriptive Performance (Training)	0.04%	99.87%	99.87%	99.99%	99.97%	99.96%	99.92%	
Predictive Performance (Validation)	0.05%	99.88%	99.88%	99.98%	99.93%	99.96%	99.91%	
Real Performance (Test)	0.05%	99.83%	99.83%	99.98%	99.95%	99.95%	99.89%	
Real Performance (External Test 1)	0.04%	99.86%	99.86%	99.98%	99.96%	99.95%	**99.91%**	
Real Performance (External Test 2)	7.63%	77.95%	77.95%	99.43%	98.53%	90.21%	**87.04%**	
BOOSTED TREE TOP-MODEL								
Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
Predictive Performance (Validation)	0.01%	99.98%	99.98%	99.99%	99.96%	99.99%	99.97%	
Real Performance (Test)	0.01%	99.99%	99.99%	99.99%	99.97%	100.00%	99.98%	
Real Performance (External Test 1)	0.02%	99.98%	99.98%	99.98%	99.95%	99.99%	**99.96%**	
Real Performance (External Test 2)	6.29%	83.63%	83.63%	98.64%	96.79%	92.49%	**89.73%**	
NEURAL NETWORK TOP-MODEL								
Descriptive Performance (Training)	0.05%	99.88%	99.88%	99.98%	99.94%	99.96%	99.91%	
Predictive Performance (Validation)	0.05%	99.86%	99.86%	99.98%	99.94%	99.95%	99.90%	
Real Performance (Test)	0.06%	99.85%	99.85%	99.97%	99.92%	99.95%	99.89%	
Real Performance (External Test 1)	0.05%	99.86%	99.86%	99.98%	99.95%	99.95%	**99.90%**	
Real Performance (External Test 2)	8.01%	79.02%	79.02%	98.34%	95.89%	90.54%	**86.64%**	