Precision Medicine

GENETIC IDENTIFICATION OF LUNG CANCER

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

UC#12 – 2024/03 (v2.1) xtractis.ai

PROBLEM DEFINITION

GOAL

Design an AI-based decision system that accurately and instantly makes a rational medical diagnosis of lung cancer from genetic sequencing of lung tissues, to determine whether it is malignant pleural mesothelioma or adenocarcinoma (ADCA).

PROS & BENEFITS

- Identify the genes involved in cancer and enhance medical knowledge by helping pulmonologists and oncologists understand the causal relationships between specific genes, their combination, and the type 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:
Gavin J. Gordon & al.,
Division of Thoracic surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Dataset:
[https://leo.ugr.es/elvira/DBCRepository/LungCancer/LungCancer-Harvard2.html]

Variable to Predict: The model diagnoses the sampled lung tissue as **ADCA** | **MESOTHELIOMA**

Potential Predictors: 12,533 variables are the level of expression of genes characterizing each patient, normalized to the median.

Observations: 213 genetic sequencing of lung tissue from patients with or without cancer.

149 cases compose a Learning Dataset for model induction using Training and Validation Datasets.

64 samples from a different experiment compose an External Test Dataset to check the top-model's performance on real unknown data and for benchmarking.

Learning Dataset: 149 patients | 69.95% 80% for Training, 20% for Validation	
ADCA	MESOTHELIOMA
134 | 90%	15 | 10%

External Test Dataset: 64 patients | 30.05%	
ADCA	MESOTHELIOMA
32 | 50%	32 | 50%

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 2 predictors among the 2 variables that XTRACTIS automatically identified as significant (out of the 12,533 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).

☑ **Ready to Deploy** 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	ADCA ⚠ MESOTHELIOMA ⚠ Automated Decision (identify cancer type)
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 v12.2.44169

1. We launch 100 inductive reasoning strategies; 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.
2. 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).
3. Among the 300 induced CVEs, the top-CVE with the best predictive performance remains complex: 206 rules share 68 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 44,700 new cases simulated by deduction from the top-CVE, around the 149 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.

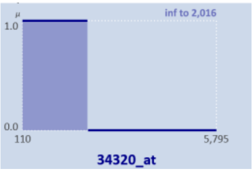
Total number of induced unitary models	Criterion for the induction optimization	Validation criterion for the top-model selection	Duration of the process (Induction Power FP64)
12,000 IVEs	F_1-Score	F_1-Score	4 days (1 Tflops)

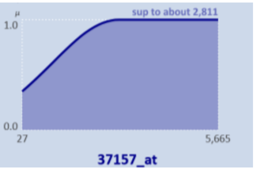
TOP-MODEL STRUCTURE

The top-IVE model has an excellent intelligibility -and is very simple- as it combines 2 predictors into only 2 rules with 1.5 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

- 2 genes identified out of 12,533
- Ranked by individual contribution (2 strong signals): #1 gene 37157_at / #2 gene 34320_at
- Labeled by binary and fuzzy classes
  Examples: **binary interval** "inf to 2,016"
  fuzzy interval "sup to about 2,811"





RULES

- 2 connective fuzzy rules without chaining
- 1 to 2 predictors per rule (on average, 1.5 predictor per rule)
- Example: fuzzy rule **R2** uses 1 predictor and concludes MESOTHELIOMA.
  Another binary rule completes this model.

IF	gene 37157_at	IS	sup to ~2,811
THEN	Diagnosis	IS	MESOTHELIOMA

Literally, the sampled lung tissue gets a mesothelioma diagnosis if the level of expression of gene #37157 is over around 2,811.

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	99.93%	100.00%	99.87%	100.00%
Classification Error	0.01%	0.00%	0.03%	0.00%

EXPLAINED PREDICTIONS FOR 3 UNKNOWN CASES

Powered by: XTRACTIS® PREDICT v12.2.44169

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

PATIENT #11

actual value
= MESOTHELIOMA

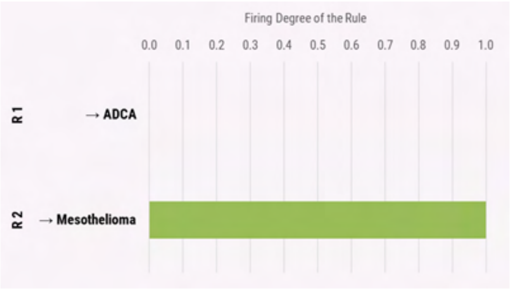
gene 34320_at	2,906
gene 37157_at	3,409

Real Time

For this patient, 1 rule is triggered:

R2 is fired at 1.000 to conclude MESOTHELIOMA.

R1 is not activated.



NUMBER OF TRIGGERED RULES
1 / 2

FUZZY PREDICTION
{ MESOTHELIOMA | 1.000 }

FINAL PREDICTION
{ MESOTHELIOMA }

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

MESOTHELIOMA ⚠

PATIENT #27

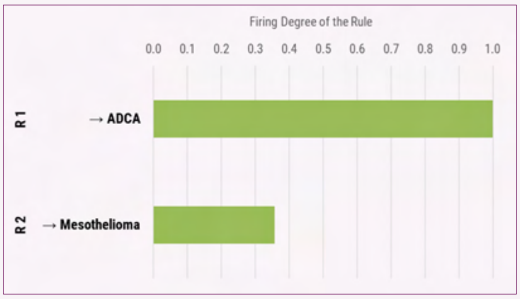
actual value
= ADCA

gene 34320_at	283
gene 37157_at	57

Real Time

For this patient, 2 rules are triggered:

R1 is fired at 1.000 to conclude ADCA,
and **R2** at 0.357 to conclude MESOTHELIOMA.



NUMBER OF TRIGGERED RULES
2 / 2

FUZZY PREDICTION
{ ADCA | 1.000,
MESOTHELIOMA | 0.357 }

FINAL PREDICTION
{ ADCA }

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

ADCA ⚠

PATIENT #15

actual value
= MESOTHELIOMA

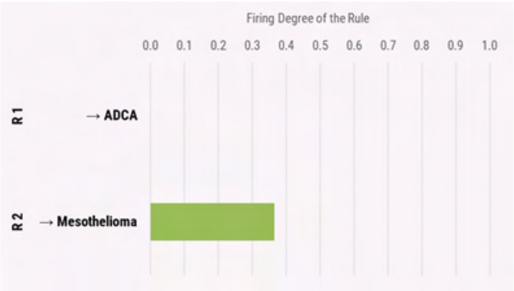
gene 34320_at	3,360
gene 37157_at	90

Real Time

For this patient, 1 rule is triggered:

R2 is fired at 0.366 to conclude MESOTHELIOMA.

R1 is not activated.



NUMBER OF TRIGGERED RULES
1 / 2

FUZZY PREDICTION
{ MESOTHELIOMA | 0.366 }

FINAL PREDICTION
{ MESOTHELIOMA }

The system delivers a correct diagnosis of the type of cancer compared to that given by the genetic oncologist, despite uncertainty (Possibility = 0.366):

MESOTHELIOMA ⚠

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

MODELING PARAMETERS

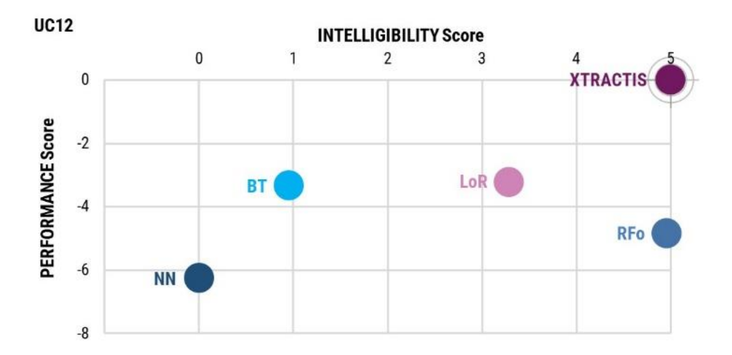
	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
MODELS RELEASE	2022/12	2023/01	2022/12	2022/12	2023/01
ALGORITHM VERSION	XTRACTIS **REVEAL** 12.2.44169	Python 3.9 | Scikit-Learn 1.1.2	Python 3.9 | LightGBM 3.3.2	Python 3.9 | LightGBM 3.3.2	Python 3.9 | TensorFlow 2.10.0 | Keras 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	20×5 folds for each CVE model	20×5 folds for each CVE model	20×5 folds for each CVE model
NUMBER OF EXPLORED STRATEGIES(1)	100 induction strategies for the CVE on Training / Validation data. 2,000 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 300 CVEs. Then Top-IVE among 2,000 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

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
NUMBER OF PREDICTORS (out of 12,533 Potential Predictors)	**2**	**7**	**7**	**6**	**12,533**
AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	**1.5** per rule	**7.0** per equation	**1.5** per rule	**1.2** per rule	**9,400.5** per equation
STRUCTURE OF THE DECISION SYSTEM	**2** disjunctive fuzzy rules without chaining Rules are not necessarily triggered at the same time to compute a prediction	**1** linear equation	**4** trees without chaining **11** binary rules	**1** chain of **5** trees **11** binary rules Tree #N corrects the error of the N-1 previous trees	**1** hidden layer | **3** hidden nodes **4** equations 3 unintelligible synthetic variables

TOP-MODEL SCORES

	Random(3)	XTRACTIS	LoR	RFo	BT	NN
INTELLIGIBILITY Score(4)		5.00	3.29	4.96	0.96	0.00
CVE Real Performance (F_1-Score) in External Test		100.00	96.77	93.33	93.33	100.00
Gap to CVE Leader in External Test		0.00	-3.23	-6.67	-6.67	0.00
IVE Real Performance (F_1-Score) in External Test	92.00	100.00	96.77	96.97	100.00	87.50
Gap to IVE Leader in External Test		0.00	-3.23	-3.03	0.00	-12.50
Average Real Performance in External Test	92.00	100.00	96.77	95.15	96.67	93.75
PERFORMANCE Score(4)		0.00	-3.23	-4.85	-3.33	-6.25



(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. XTRACTIS's perfect results 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; 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):

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

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):

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

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):

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

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):

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

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):

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

Intelligibility Score of T_i

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

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 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)	*11.76%*	*0.698*					***92.00%***	
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%	1 (3.13%)
IVE - Descriptive Performance (Training)	0.01%	99.99%	100.00%	99.99%	99.87%	100.00%	99.93%	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)	0.03%	99.97%	100.00%	99.97%	99.73%	100.00%	99.87%	0 (0.00%)
IVE - Real Performance (149 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)	0.00%	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)	3.13%	93.75%	93.75%	100.00%	100.00%	94.12%	96.77%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	3.13%	93.75%	93.75%	100.00%	100.00%	94.12%	96.77%	
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)	6.25%	87.50%	87.50%	100.00%	100.00%	88.89%	93.33%	
IVE - Descriptive Performance (Training)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
IVE - Real Performance (External Test)	3.13%	93.75%	100.00%	93.75%	94.12%	100.00%	96.97%	
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)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
CVE - Real Performance (External Test)	6.25%	87.50%	87.50%	100.00%	100.00%	88.89%	93.33%	
IVE - Descriptive Performance (Training)	0.67%	99.25%	100.00%	99.25%	93.75%	100.00%	96.77%	
IVE - Real Performance (External Test)	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	
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)	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)	5.37%	86.67%	86.67%	95.52%	68.42%	98.46%	76.47%	
IVE - Real Performance (External Test)	12.50%	87.50%	87.50%	87.50%	87.50%	87.50%	87.50%	