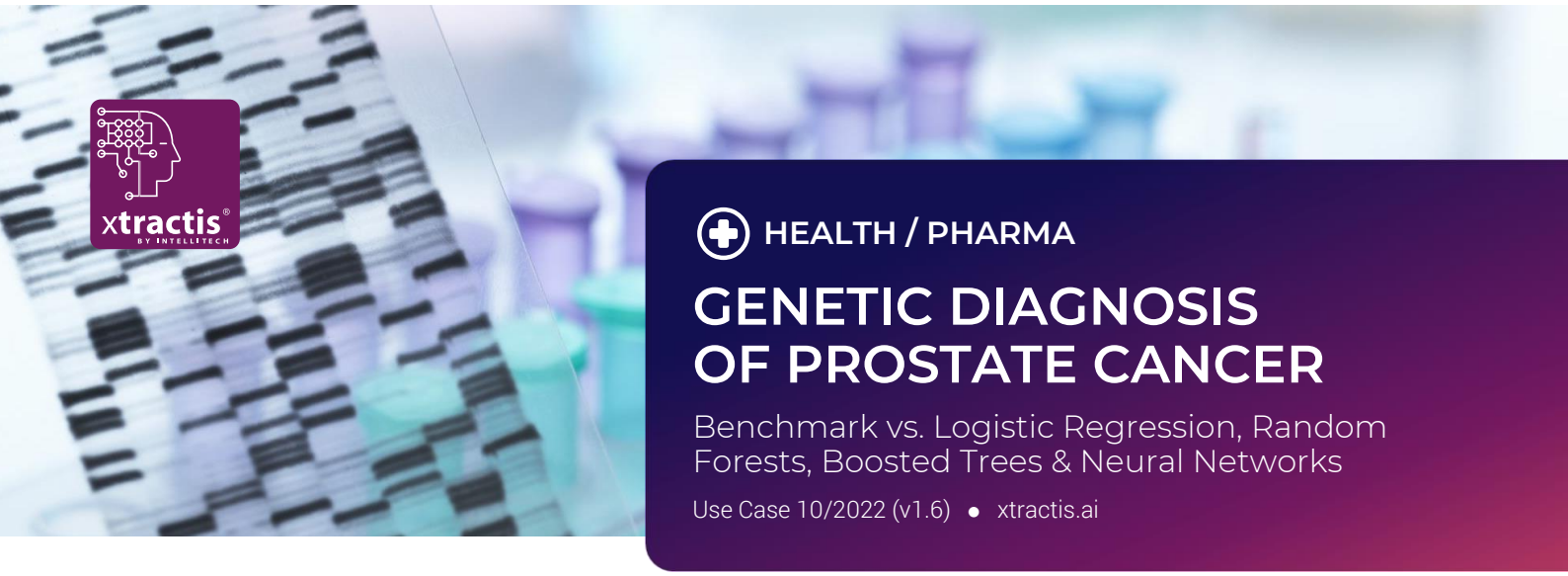


HEALTH / PHARMA

GENETIC DIAGNOSIS OF PROSTATE CANCER

Benchmark vs. Logistic Regression, Random Forests, Boosted Trees & Neural Networks

Use Case 10/2022 (v1.6) • xtractis.ai

PROBLEM DEFINITION

PROBLEM

How to make an automated —yet totally transparent— medical diagnosis of prostate cancer from genetic sequencing of prostate tissue?

GOALS & BENEFITS

- ☑ Identify the genes involved in cancer and enhance medical knowledge by helping urologists and oncologists understand the causal relationships between specific genes, 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

- ▶ **Observations**: 136 genetic sequencing of prostate tissue from patients with or without cancer, divided into 102 cases for Training/Validation and 34 cases for External Test, from a different experiment.
  Source: D. Singh & al., Department of Adult Oncology, Brigham and Women's Hospital, Harvard Medical School. www-genome.wi.mit.edu/mpr/prostate (2014)
- ▶ **Predictive Variables**: 12,600 Potential Predictors are the level of expression of genes characterizing each patient, normalized to the median.
- ▶ **Variable To Predict**: Sampled prostate tissue diagnosis [NORMAL / TUMOR].

MODEL TYPE

Regression | Multinomial Classification | **Binomial Classification** | Scoring

XTRACTIS SOLUTION

PROCESS

Reference Data → **INDUCTION + Reverse-Engineering**[1] → XTRACTIS Top-Model

New Cases → **DEDUCTION** → Automated Decision (supporting experts' diagnosis): NORMAL / TUMOR

SOFTWARE ROBOTS

XTRACTIS® GENERATE

XTRACTIS® PREDICT + Prediction Report (for decision explainability)

RESULTS

- ☑ **Intelligible Predictive Top-Model**: Decision system composed of 4 unchained gradual rules, each rule using some of the 7 variables that XTRACTIS identified as predictors.
- ☑ **Robust Predictive Top-Model**: Excellent performance on External Test.
- ☑ **Operational Efficient System**: Real-time predictions up to 70,000 decisions/s., offline or online (API).

TOP-MODEL INDUCTION

INDUCTION PARAMETERS

We launch 100 inductive reasoning strategies; each strategy is applied to 40 different 5-fold-partitions of the Training/Validation dataset to get a reliable assessment of the descriptive and predictive performances. 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 300 CVE, the top-CVE with the best predictive performance remains complex (471 predictors shared by 658 rules).

We then apply 2,000 induction strategies to the same single Training (70%)/Validation (15%)/Test (15%) partition of a synthetic dataset: 20,400 new cases simulated by deduction from the top-CVE, around the 102 cases but distinct from these original cases. This XTRACTIS Reverse-Engineering[1] process induces 2,000 IVE. The top-IVE selected is as efficient as the top-CVE, but intelligible (7 predictors shared by 4 rules).

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)
22,000 IVE	F_1-Score	F_1-Score	17 days (1 Tflops)

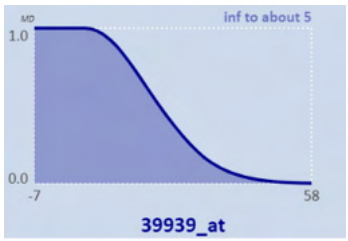
STRUCTURE

Intelligibility

The top-IVE model combines the 7 predictors automatically selected by XTRACTIS into 4 rules. Its Structure Report reveals all the internal decision logic and ensures that the human expert understands the model. This decision system is a *white-box* model that can be audited by the domain expert and certified by the regulator before its deployment to end-users.

PREDICTORS

- 7 genes identified out of 12,600
- Ranked by impact significance (2 strong, 3 medium & 2 weak signals): #1 gene 36883_at / #2 gene 37639_at / #3 /… / #7
- Labeled by fuzzy classes
  Example: **fuzzy interval** "inferior to about 5"



RULES

- 4 connective fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules)
- 2 to 4 predictors per rule (on average, 3 predictors per rule)
- Example: **fuzzy rule R4** uses 4 predictors, and concludes "TUMOR".
  3 other fuzzy rules complete this model.

IF	gene 39939_at	IS	inferior to about 5
AND	gene 35178_at	IS	inferior to about -2
AND	gene 36883_at	IS	inferior to about 87
AND	gene 40282_s_at	IS	inferior to about 77
THEN	Diagnosis	IS	TUMOR

PERFORMANCE

Robustness

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.

Performance Dataset	Synthetic Data: DESCRIPTIVE Training (70%)	Synthetic Data: PREDICTIVE Validation (15%)	Synthetic Data: REAL Test (15%)	Reference Data: REAL External Test
F_1-Score	99.36%	99.52%	99.86%	100.00%
Classification Error	0.65%	0.49%	0.36%	0.00%

Xtractis Top-Model: Intelligible AND High Predictive Capacity

PREDICTIONS FOR 3 CASES FROM THE EXTERNAL TEST SET

CASE
(not used in Training/Validation)

DEDUCTIVE INFERENCE OF RULES

DECISION

PATIENT #1
(actual value = TUMOR)

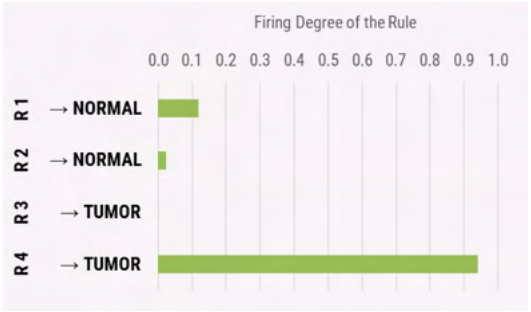
gene 39939_at	5
gene 33792_at*	1.7
gene 35178_at	2
gene 36883_at	39
gene 37639_at	162
gene 37367_at	114
gene 40282_s_at	26

Real Time

For this patient, 3 rules are triggered:

R4 is fired at 0.940, **R1** at 0.117, and **R2** at 0.022.

R3 is not activated.



NUMBER OF TRIGGERED RULES
3 / 4

FUZZY PREDICTION
{ TUMOR|0.940, NORMAL|0.117 }

FINAL PREDICTION
{ TUMOR }

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

TUMOR

PATIENT #30
(actual value = NORMAL)

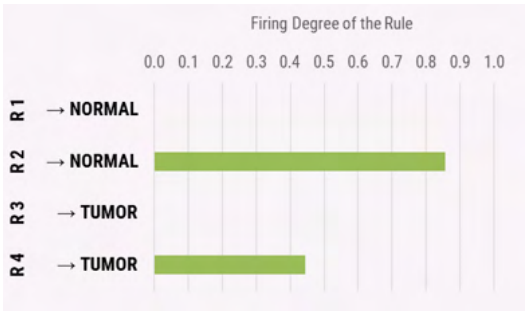
gene 39939_at	24
gene 33792_at	296.9
gene 35178_at	2
gene 36883_at	21
gene 37639_at	33
gene 37367_at	92
gene 40282_s_at	60

Real Time

For this patient, 2 rules are triggered:

R2 is fired at 0.857, and **R4** at 0.445.

R1 and R3 are not activated



NUMBER OF TRIGGERED RULES
2 / 4

FUZZY PREDICTION
{ NORMAL|0.857, TUMOR|0.445 }

FINAL PREDICTION
{ NORMAL }

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

NORMAL

PATIENT #5
(actual value = TUMOR)

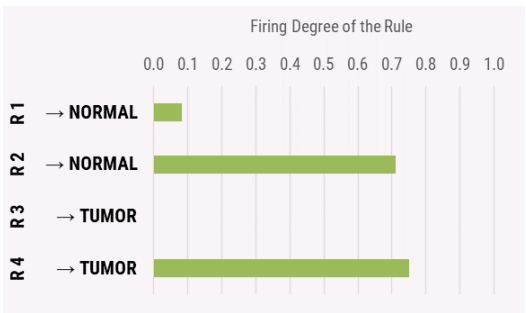
gene 39939_at	14
gene 33792_at	20.6
gene 35178_at	4
gene 36883_at	20
gene 37639_at	55
gene 37367_at	75
gene 40282_s_at	46

Real Time

For this patient, 3 rules are triggered:

R4 is fired at 0.751, **R2** at 0.711, and **R1** at 0.082.

R3 is not activated.



NUMBER OF TRIGGERED RULES
3 / 4

FUZZY PREDICTION
{ TUMOR|0.751, NORMAL|0.711 }

FINAL PREDICTION
{ TUMOR }

The system delivers a correct diagnosis of cancer compared to that given by the genetic oncologist, despite uncertainty/hesitation:

TUMOR

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-IVE BENCHMARK

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREES	NEURAL NETWORK
MODELS RELEASE	2021/06	2022/10	2021/08	2021/04	2022/03
ALGO VERSION	XTRACTIS **GENERATE** 11.2.38531	Python 3.9.12, 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 (for the reverse engineering of top-CVE): 70% Training; 15% Validation; 15% 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[2]	100 induction strategies for the CVE on Training / Validation data 2,000 induction strategies for the IVE on simulated 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
NUMBER OF MODELS	300 CVE + selection of the top-CVE 2,000 IVE (for the reverse engineering of top-CVE) + selection of the top-IVE	300 CVE + selection of the top-CVE 1 top-IVE	300 CVE + selection of the top-CVE 1 top-IVE	300 CVE + selection of the top-CVE 1 top-IVE	300 CVE + selection of the top-CVE 1 top-IVE
TOP-IVE STRUCTURE					
NUMBER OF PREDICTORS (out of 12,600 Potential Predictors)	7	120	19	24	12,600
DECISION STRUCTURE	System with 4 unchained fuzzy rules (or 2 disjunctive fuzzy rules)	1 linear equation	15 trees; 50 binary rules	14 chained trees; 48 binary rules	1 hidden layer; 13 hidden nodes
MODEL INTELLIGIBILITY (& DECISION EXPLAINABILITY)	⊕⊕⊕ 3 predictors per rule on average; only a few rules are triggered at a time	⊖○○ Linear equation with 120 coefficients	⊖○○ Lots of predictors and rules	⊖⊖○ Tree #N corrects the error of the N-1 previous trees	⊖⊖⊖ Unintelligible synthetic variables
TOP-IVE REAL PERFORMANCE (External Test)					
Classification Error *Random*[3] *11.76%*	**0.00%**	**8.82%**	**29.41%**	**20.58%**	**2.94%**
Sensitivity	100.00%	96.00%	92.00%	92.00%	96.00%
Specificity	100.00%	77.78%	11.11%	44.44%	100.00%
PPV	100.00%	92.31%	74.19%	82.14%	100.00%
NPV	100.00%	87.50%	33.33%	66.67%	90.00%
F_1-Score *92.00%*	**100.00%**	**94.11%**	**82.14%**	**86.79%**	**97.96%**
Refusals N/A	0.00%	N/A	N/A	N/A	N/A
MODEL ROBUSTNESS[4]	#1	#3	#5	#4	#2

[1] Given the small number of reference cases of this dataset, the XTRACTIS Reverse-Engineering (CVE→IVE) is necessary to get a robust AND intelligible model.

[2] All CVE and IVE models are optimized according to their validation F_1-Score. The XTRACTIS top-CVE and top-IVE are selected according to their validation F_1-Score while checking that it remains close to their training F_1-Score. The ML/LoR CVE top-models are selected according to their F_1-Score mean value in validation. Each ML/LoR top-IVE is obtained by applying the respective ML/LoR top-CVE strategy on 100% of the Training/Validation data.

[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).

[4] The perfect results of the XTRACTIS 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/