HEALTH / PHARMA

GENETIC IDENTIFICATION OF LUNG CANCER

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

Use Case 2023/01 (v1.1) • xtractis.ai

PROBLEM DEFINITION

PROBLEM

How to make an automated — yet totally transparent — medical diagnosis of lung cancer from genetic sequencing of different tissues?

GOALS & 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

- ▶ **Observations**: 149 genetic sequencing of lung tissue from patients with Mesothelioma (15 | 10%) or ADCA (134 | 90%) cancers, for Training/Validation, and 64 samples for External Test, from a different experiment (Mesothelioma 32 | 50%, ADCA 32 | 50%).
  Source: Gavin J. Gordon & al., Division of Thoracic surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts [https://leo.ugr.es/elvira/DBCRepository/LungCancer/LungCancer-Harvard2.html]
- ▶ **Predictive Variables**: 12,533 Potential Predictors are the level of expression of genes characterizing each patient, normalized to the median.
- ▶ **Variable To Predict**: lung cancer diagnosis [MESOTHELIOMA / ADCA].

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): ADCA | MESOTHELIOMA

SOFTWARE ROBOTS

XTRACTIS® **GENERATE**

XTRACTIS® **PREDICT** + **PREDICT**ion Report (for decision explainability)

RESULTS

- ☑ **Intelligible Predictive Top-Model**: Decision system composed of 2 unchained gradual rules, each rule using one or two variables that XTRACTIS identified as predictors.
- ☑ **Robust Predictive Top-Model**: Perfect Real Performance in 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 20 different 5-fold-partitions of the Training/Validation dataset to get a reliable assessment of the descriptive and predictive performances. 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)**. Among the 300 CVE, the top-CVE with the best predictive performance remains complex (68 predictors shared by 206 rules).

We then apply 2,000 induction strategies to the same single Training (34%)/Validation (33%)/Test (33%) partition of a synthetic dataset: 44,700 new cases simulated by deduction from the top-CVE, around the 149 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 (2 predictors shared by 2 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)
12,000 IVE	F_1-Score	F_1-Score	17 days (1 Tflops)

STRUCTURE

Intelligibility

The top-IVE model is very simple as it combines the 2 predictors automatically selected by XTRACTIS into 2 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

- 2 genes identified out of 12,533
- Their impact significance (2 strong signals):
  #1 gene 37157_at / #2 gene 34320_at
- Labeled by 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 predictors per rule)
- Example: **fuzzy rule R2** uses 1 predictor and concludes "Mesothelioma".
  An other binary rule completes this model.

IF	gene 37157_at	IS	sup to about 2,811
THEN	Diagnosis	IS	Mesothelioma

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.

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	Synthetic Data			Reference Data
Dataset	DESCRIPTIVE 34% Training	PREDICTIVE 33% Validation	REAL 33% Test	REAL External Test
F_1-Score	99.93%	100.00%	99.87%	100.00%
Classification Error	0.01%	0.00%	0.03%	0.00%

Xtractis Top-Model: Intelligible AND High Predictive Capacity

EXPLAINED PREDICTIONS FOR 3 CASES FROM THE EXTERNAL TEST SET

CASE
(not used in Training/Validation)

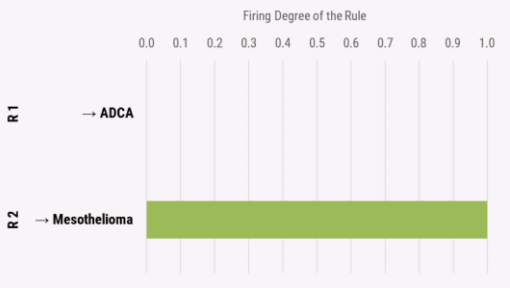
DEDUCTIVE INFERENCE OF RULES

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.
R1 is not activated.



NUMBER OF TRIGGERED RULES
1 / 2

FUZZY PREDICTION
{ Mesothelioma | 1.000 }

FINAL PREDICTION
{ Mesothelioma }

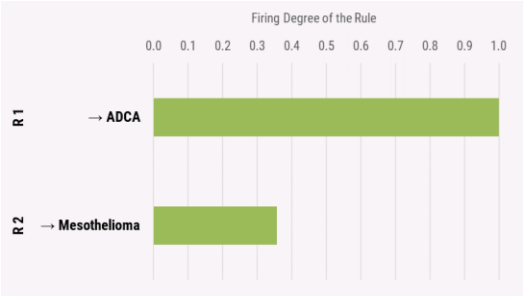
The system delivers a correct diagnosis of the type of cancer compared to that given by the genetic oncologist:

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, and **R2** at 0.357.



NUMBER OF TRIGGERED RULES
2 / 2

FUZZY PREDICTION
{ ADCA |1.000,
Mesothelioma |0.357 }

FINAL PREDICTION
{ ADCA }

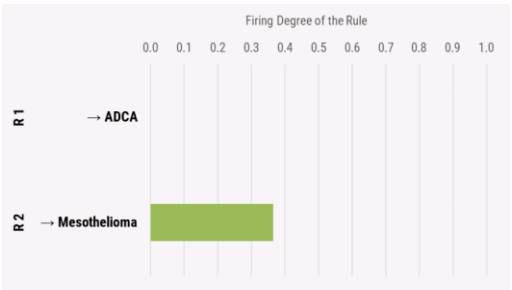
The system delivers a correct diagnosis of the type of cancer compared to that given by the genetic oncologist:

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.
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-IVE BENCHMARK

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREES	NEURAL NETWORK
MODELS RELEASE	2022/12	2023/01	2022/12	2022/12	2023/01
ALGO VERSION	XTRACTIS **GENERATE** 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 (for the reverse engineering of top-CVE): 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[2]	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
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 12,533 Potential Predictors)	**2**	**7**	**7**	**6**	**12,533**
DECISION STRUCTURE	System with **2** unchained fuzzy rules	**1** linear equation with 7 coefficients.	**4** trees; **11** binary rules	**5** chained trees; **11** binary rules	**1** hidden layer; **3** hidden nodes
MODEL INTELLIGIBILITY (& DECISION EXPLAINABILITY)	⊕⊕⊕ Rules not necessarily triggered at the same time	⊕⊕○ A few predictors and coefficients	⊕○○ A few predictors and more rules	⊖⊖○ Tree #N corrects the error of the N-1 previous trees	⊖⊖⊖ Unintelligible synthetic variables
TOP-IVE REAL PERFORMANCE (External Test)					
	Random[3]				
Classification Error	*11.76%*	**0.00%**	**3.13%**	**3.13%**	**0.00%**	**12.50%**
Sensitivity		100.00%	93.75%	100.00%	100.00%	87.50%
Specificity		100.00%	100.00%	93.75%	100.00%	87.50%
PPV		100.00%	100.00%	94.12%	100.00%	87.50%
NPV		100.00%	94.12%	100.00%	100.00%	87.50%
F_1-Score	*92.00%*	**100.00%**	**96.77%**	**96.97%**	**100.00%**	**87.50%**
Refusals	N/A	0.00%	N/A	N/A	N/A	N/A
MODEL ROBUSTNESS[4]		#1	#4	#3	#1	#5

[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 and BT Top-IVE on External Test could be explained by a low number of reference points compared to the very large number of potential predictors. Conversely, NN needs much more reference points to deliver a good performance on External Test.

More Use Cases:
xtractis.ai/use-cases/