LUCIA

LUng Cancer-related risk factors and their Impact Assessment

HORIZON-MISS-2021-CANCER-02

LUCIA Workshop – Understanding Lung Cancer

San Sebastian, Sept. 5th, 2023

Iván Macía
VICOMTECH

Kiko Núñez
SAS-FISEVI

European Commission

Mathematical & Computational Models for Risk Estimation and Disease Understanding

Iván Macía, PhD

imacia@vicomtech.org

Director of Digital Health & Biomedical Technologies

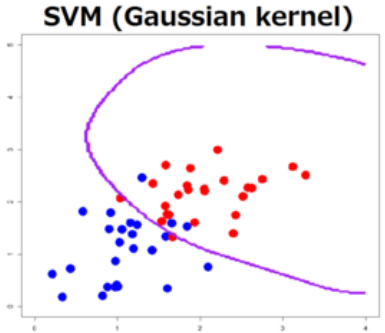
vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

According to their mathematical nature

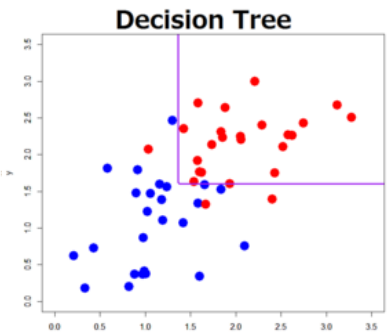
- **Knowledge-based**: based on a explicit modelling of **knowledge** or **experience**.
  - The most simple models use **rules**, possibly with **ontology** support, whereas more advanced models use structures such as **knowledge graphs**, allowing computations
  - *Example*: implementing **rule-based protocols** or guidelines for LC screening/diagnosis/treatment

- **Statistical Models**: adjust parameters of a statistical parametrical function to fit some multivariable data for **classification** or **regression**.
  - Fitting is usually done by optimization, which limits the size of dataset / number of variables
  - *Example*: statistical models (e.g. multivariate linear/Cox regression, logistic regression) for based in observational studies with large cohorts where risk factor (RF) data is collected, usually through questionnaires. Some of these models are simplified into calculators, for risk prediction and easy adoption

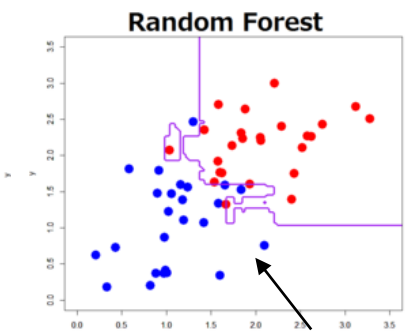
- **Machine Learning Models**: develop mathematical representations of data **by iteratively adjusting (training)** the parameters of a complex function for classification or regression.
  - The simplest are similar to statistical models, but allow for more complex models representing **non-linear** functions, as well as **N input dimensions** and **M output classes**
  - Able to handle a **large number of input variables (features)** without a priori hypothesis. Unstructured data can only be handled by hard pre-processing work (feature extraction)
  - *Examples*: classifiers/regressors such as **decision trees, support vector machines, random forests, boosting**… for risk estimation or diagnostics of LC



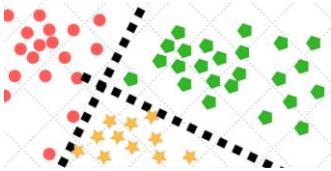
Two-variable, two-class classification



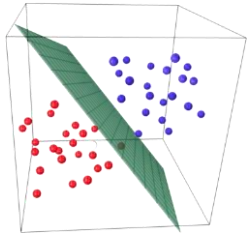
Non-linearity



Overfitting



3 classes

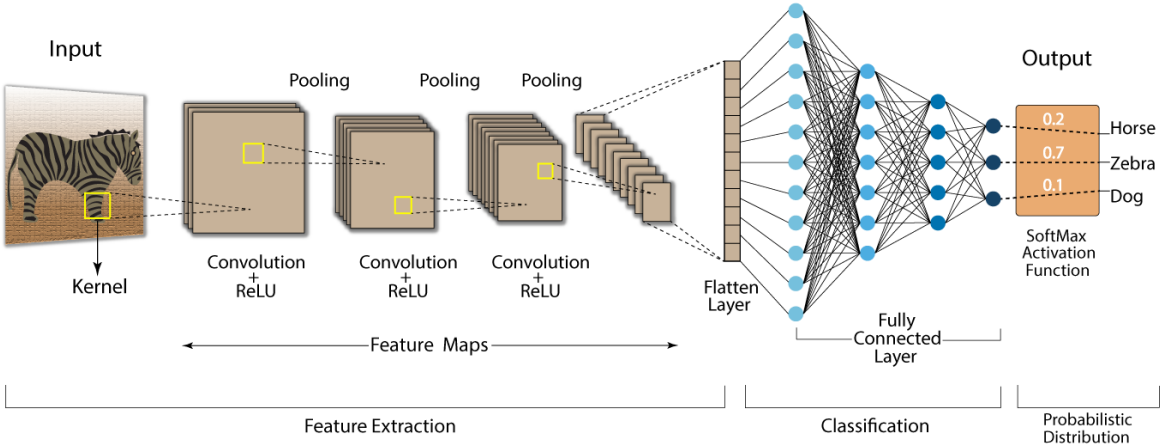


3 input variables / features

LUCIA Modelling Overview – Types of Models

- **Deep Learning Models**: advanced ML models based on **last generation neural networks**.
  - They are exceptionally good in dealing with **unstructured data** or large complex data, such as **images, natural language, signals** or **time series**. They have the ability to extract the relevant features / patterns of the data and perform classification / prediction **end-to-end**, without costly and biased pre-processing. They can handle naturally heterogeneous data
  - *Examples*: estimate risk or LC or diagnosis by a convolutional neural network on CT images, encode temporal diagnosis / treatments in the EHR for prediction

End-to-end image classification without prior image analysis by Deep Learning Convolutional Neural Network (DL-CNN)



LUCIA Data and Model Landscape for Understanding LC

Lifestyle & Exposure

Discover / assess RFs in populations and predict risk

Data sources:
- Electronic Health Records
- Linked open data
- Questionnaires and apps
- Sensor data

Models / Analysis:
- Statistical/ML models on RFs
- ML/DL EHR models for incident prediction

SAMPLE SIZE ★★★* (populations)
DATA DEPTH ★★
QUALITY ★

*But follows poulation prevalence of cases (low) vs controls (very high)

Risk Factor Analysis

Detailed analysis of RFs, causality and transformation potential linking RFs to biology

Data sources:
- Cohorts with RFs
- Bibliography and evidence on specific RFs
- DBs e.g. chemical, molecular…

Models / Analysis:
- Knowledge models
- GeoAI models
- Integrative models

SAMPLE SIZE ★★ (observ cohorts)
DATA DEPTH ★
QUALITY ★★ (questionnaires)

Biology

Understand molecular, celullar, immune changes leading to disease

Data sources:
- Digital data and biobank samples (cohorts/RWD)
- Open research data
- Omic DBs (e.g. variants)

Models / Analysis:
- GWAS, variant analysis
- Polygenic risk scores
- Integrative omic models
- Systems biology models

SAMPLE SIZE ★ (expensive)
DATA DEPTH ★★
QUALITY ★★★

Disease

In-depth study disease phenotypes and develop early / precision diagnostics

Data sources:
- Clinical data & demographics
- Cohorts & biobank samples
- Imaging (MR…) & Pathology
- Omics data & sensor data
- Clinical guidelines

Models / Analysis:
- Analysis of Real World Data
- Imaging/omics/sensor AI
- Deep phenotyping AI models

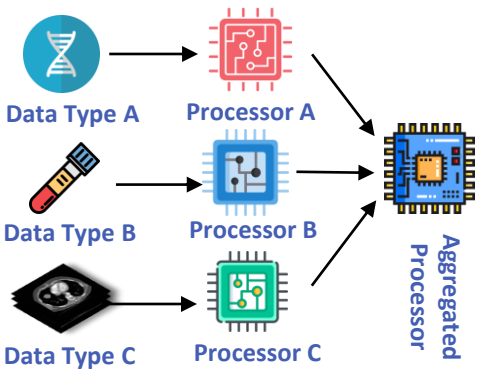
SAMPLE SIZE ★★ (real world data)
DATA DEPTH ★★★
QUALITY ★☆ (depends on DB)

Integrating Heterogeneous Data and Models

LUCIA
LUng Cancer-related risk factors and their Impact Assessment

EARLY INTEGRATION / FUSION

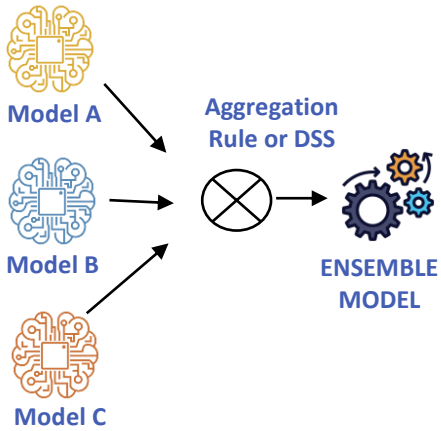
MODELS TRAINED WITH HETEROGENEOUS DATA



Models integrate heterogeneous data within their structure

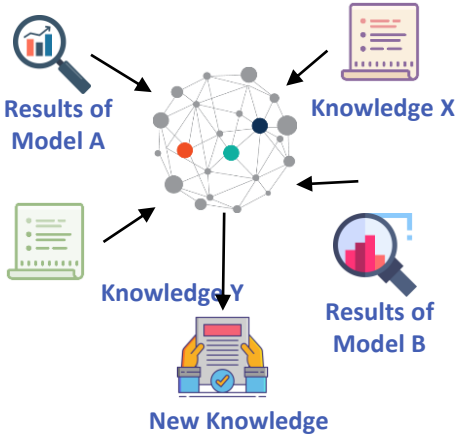
LATE INTEGRATION / FUSION

ENSEMBLE MODELS



Model outputs are aggregated e.g. using rules or other (ML) models

INTEGRATED KNOWLEDGE MODELS



Data-driven model insights and existing knowledge are integrated for reasoning

Lung Cancer Predictive Models in LUCIA
(Retrospective Study)

Kiko Núñez

SAS-FISEVI

fjose.nunez@juntadeandalucia.es

Hospital Universitario Virgen Macarena

FISEVI
Fundación Pública Andaluza para la Gestión de la Investigación en Salud de Sevilla

TALK OUTLINE:

- LC-related risk factors reported in the literature
- Predictive models for LC risk
- LUCIA approach to LC risk modelling

LC-related risk factors reported in the literature

General population

Non-smokers

Social disparities

CHALLENGES FOR DEVELOPING LC RISK MODELS

- Heterogeneous domains (clinical, genetic, lifestyle, occupational, environmental, and social)
- A mix of longitudinal, static, time-dependent, seasonal, etc… variables with potential interrelationships, correlations and biases
- Available retrospective (real-world) data is usually non-standardized and low quality for research purposes

Estrogen

Eur Respir J 2016; 48: 889–902

chloride)

Air pollution

Estrogen

Curr Probl Cancer 2017; 41(5): 328–339

Thorac Surg Clin. 2022; 32(1): 23-31

Predictive models for LC risk

$PLCO_{M2012}$

Context: Secondary analysis of NLST data
Model: Logistic regression
Population: US 22.229 (NLST + ACRIN)
Risk factors:

- Screening results based on Lung-RADS (LDCT): +/- (3 in 1 year)

Endpoint: Stratified risk in 1-4 years
Outcomes: AUC: 0.761

JAMA Netw Open 2019; 2(3):e190204

LLP_{V2}

Context: Case-control study
Model: Multivariable logistic regression
Population: UK 579 cases + 1157 age- and sex- matched controls
Risk factors:

- Smoking duration
- Prior dx of pneumonia
- Occupational exp. to asbestos
- Prior dx of malignant tumor
- Family history of lung cancer

Endpoint: Absolute risk in 5 years
Outcomes: AUC: 0.71

Br J Cancer. 2008; 98(2): 270–276

Considerations

Context: Applied in US and UK only
Models: Classic statistics, not AI-based
Population: Non-smokers not included. Mostly white/Caucasian ethnic
Risk factors: Missing genetic, environmental and social determinants

LUCIA approach to LC risk modelling

Phase I

Context: EHR-based retrospective RWD
Model: Benchmark of classic models + ML/DL models
Population: ES+BE+LV 0.5M (est.)
1:10 cases-controls ratio
(Candidates) risk factors:
- Clinical (inpatient, outpatient, lab, prescriptions, history)

Endpoint: Absolute risk in 1 year
Outcomes: AUC, RMSE, R^2

Phase II

Context: Phase I + Geo-referenced open data
Model: Benchmark of classic models + ML/DL models
Population: ES+BE+LV 0.5M (est.)
1:10 cases-controls ratio
(Candidates) risk factors: Phase I +
- Lifestyle (alcohol, smoking, BMI)
- Environmental (pollution, radon)
- Social determinants

Endpoint: Absolute risk in 1 year
Outcomes: AUC, RMSE, R^2

Challenges

Context: RWD sources (curation and preprocessing)
Models: Federated infrastructure GDPR-compliant
Population: Biases related to unbalance of samples between different data providers
Risk factors: Link outcomes of this study with deep phenotyping study