



Artificial Intelligence for Case-Finding in Celiac Disease: Lessons from the NEEDLE Study

Salvatore Oliva

Professore Associato di Pediatria

Dipartimento Materno Infantile e Scienze Urologiche, Sapienza – Università di Roma

UOSD Fibrosi Cistica e Gastroenterologia Pediatrica

Azienda Ospedaliero Universitaria Policlinico Umberto I



SAPIENZA
UNIVERSITÀ DI ROMA

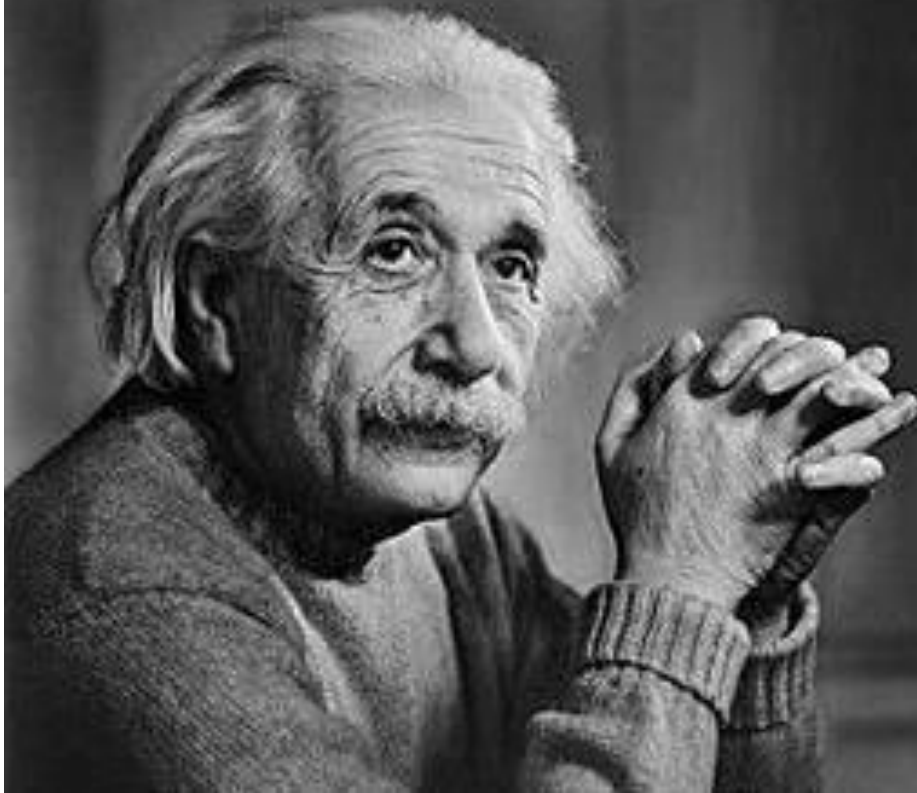


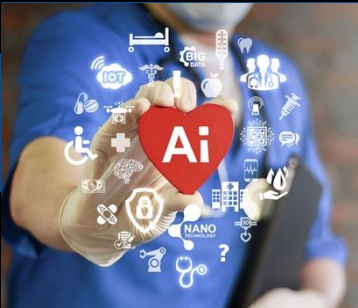
UMBERTO I
POLICLINICO DI ROMA

I'm not a bioinformatic

If you can't explain it **simply**, you
don't understand it well enough.

– Albert Einstein





What's AI?

AI stands for artificial intelligence, which is the **simulation of human intelligence** processes by machines or computer systems.

AI can mimic **human capabilities** such as **communication**, **learning from experience**, and **decision-making**

What's AI?

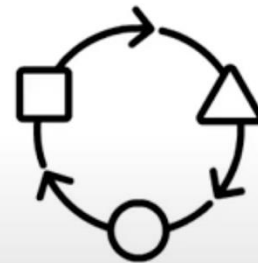
☒ Generalized learning

☒ Reasoning

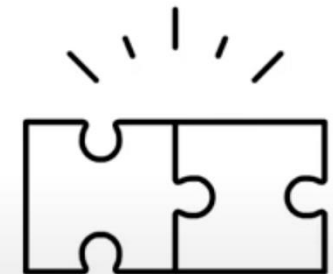
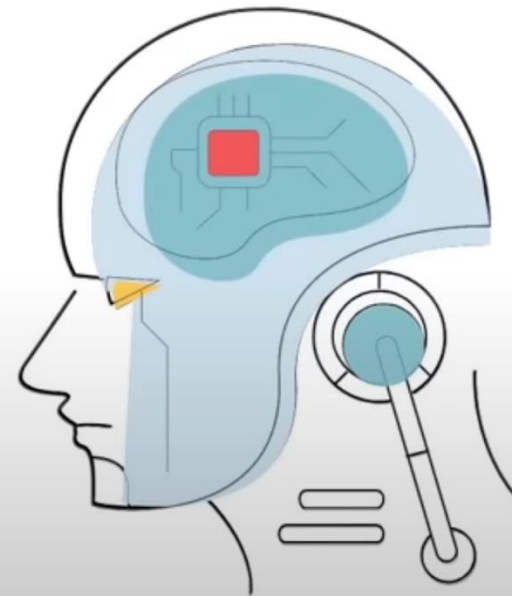
☒ Problem solving



REASON



ADAPT

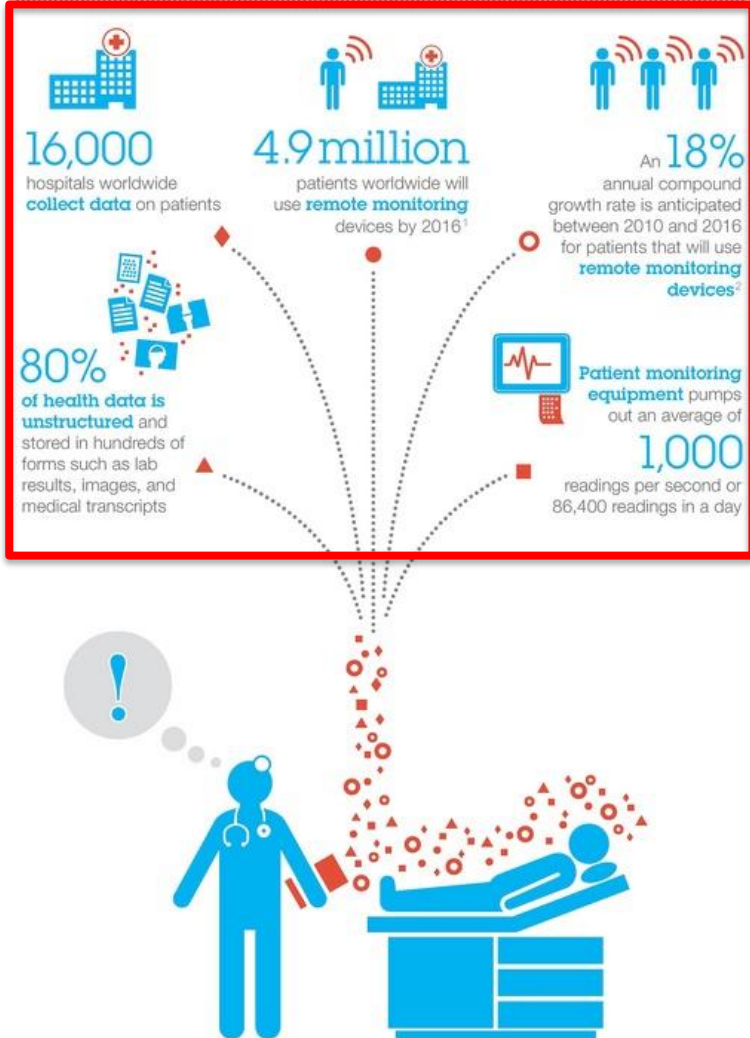


PROVIDE SOLUTIONS



Big Data in Healthcare: Tapping New Insight to Save Lives

Healthcare is challenged by large amounts of data in motion that is diverse, unstructured and growing exponentially. Data constantly streams in through interconnected sensors, monitors and instruments in real-time faster than a physician or nurse can keep up.



As the volume and velocity of health data increases, new technologies such as **Stream Computing** that analyzes health information in real-time and **big data analytics** that can predict the on-set of illness can be used to help caregivers make better decisions.

AI and **BIG DATA**

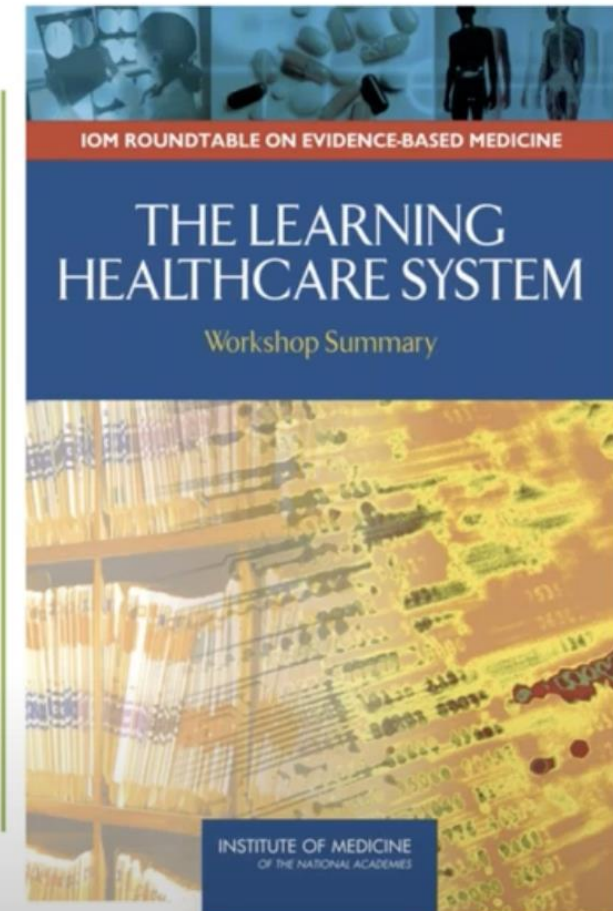
- Large amount of patient data stored in EHRs
- **Widespread digitalization** in healthcare (exponentially increasing of data from many different sources)
- Potential rich source for **research and data mining**

AI and **BIG DATA**

“Professional judgement will be always
vital to shaping care,

but ...

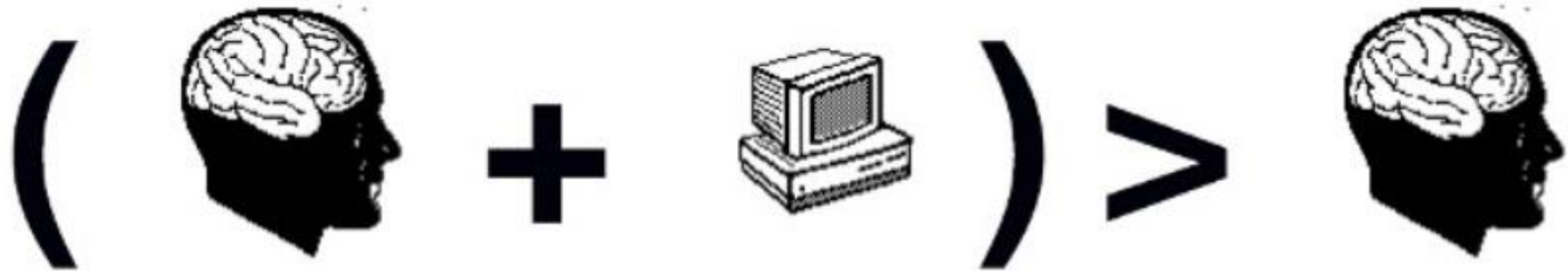
the amount of *information for any given
decision is moving beyond* **unassisted
Human Capacity**”



Olsen, LeighAnne, Dara Aisner, and J. Michael McGinnis.
“The learning healthcare system.” (2007).

Fundamental theory of informatics

Goal of informatics is



Goal is not





NEEDLE

**MACHINE LEARNING
AS A NEW METHOD FOR
“CASE E -FINDING” IN
CELIAC DISEASE E**



Received: 6 June 2024

Accepted: 24 February 2026

DOI: 10.1002/jpn3.70446

JPGN

ORIGINAL ARTICLE

Gastroenterology: Celiac Disease

The needle study: Machine learning as a new method for case-finding in celiac disease

Chiara Maria Trovato¹ | Monica Montuori² | Maria Ludovica Costanzo³ |
Giusy Russo² | Cosimo Ruggiero² | Danila Volpe² | Francesca Ferretti¹ |
Antonella Diamanti¹ | Carlo Ciliberto⁴ | Salvatore Oliva² 



SAPIENZA
UNIVERSITÀ DI ROMA

2. WHY THIS STUDY?

Celiac disease: a hidden burden



High prevalence

Celiac disease affects **>1%** of the global population (~**80 million** people worldwide)



Underdiagnosed

Up to **75%** of people with CeD remain undiagnosed



Clinical variability

From **typical** gastrointestinal symptoms to **atypical** extra-intestinal manifestations



Diagnostic challenge

Current strategies rely on clinical suspicion and serology, **missing many cases**



Opportunity with AI

Machine learning can uncover non-obvious patterns in routine clinical data to **guide screening**

THE CELIAC DISEASE ICEBERG

Diagnosed

~25%

Visible tip

Patients with typical symptoms or positive serology



Undiagnosed

~75%

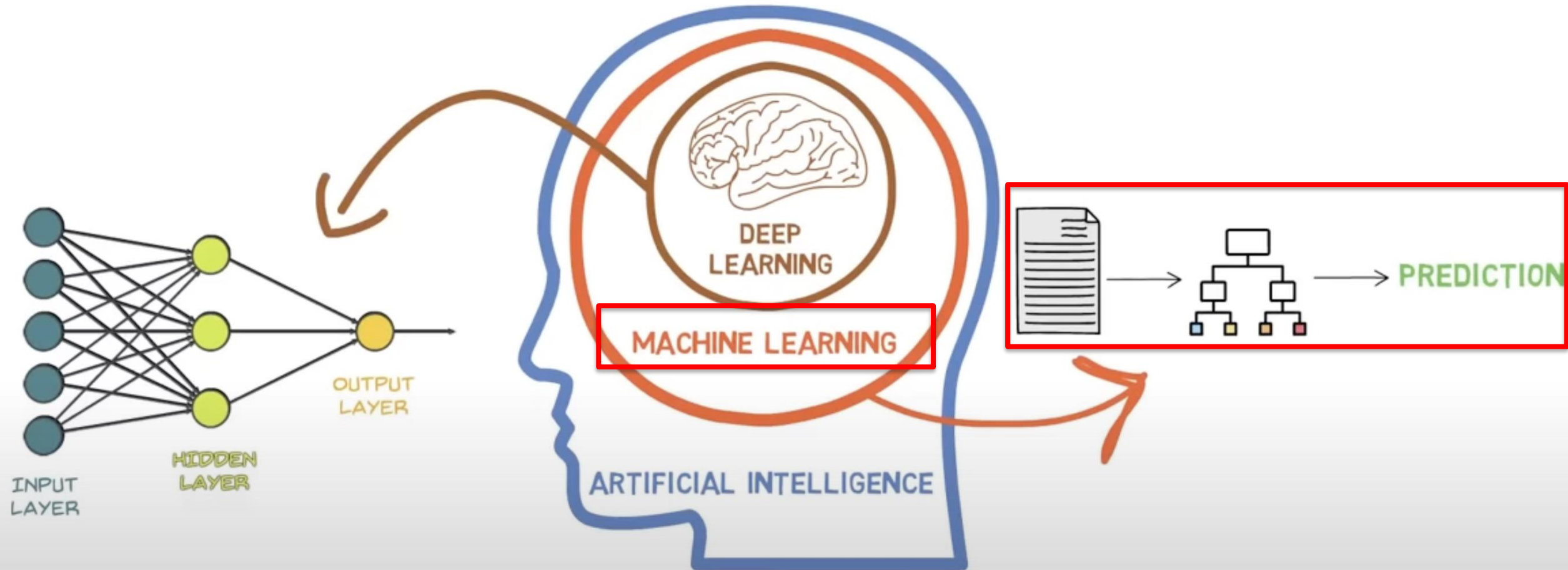
Hidden mass

Patients with non-specific or atypical symptoms not identified by current approaches

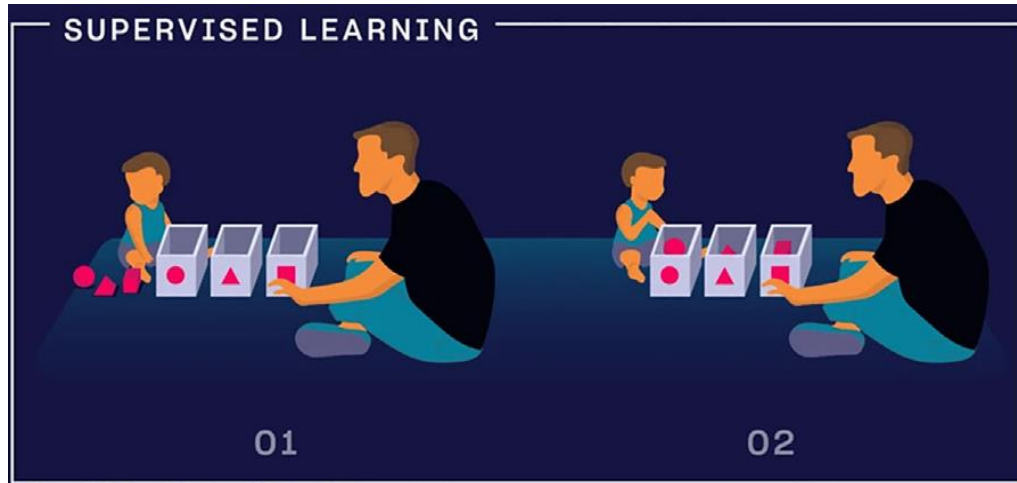


We need better tools to identify children who should be screened before celiac disease remains hidden.

How AI works



Learning process: supervised vs unsupervised

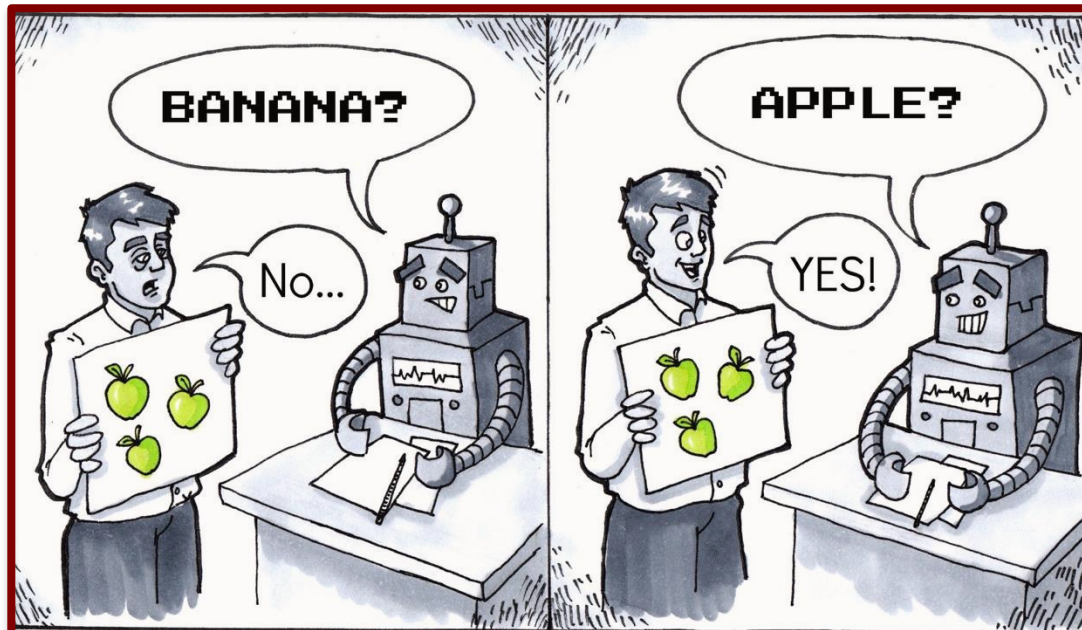


The teacher (developer) knows what they want **to teach the child** (AI)

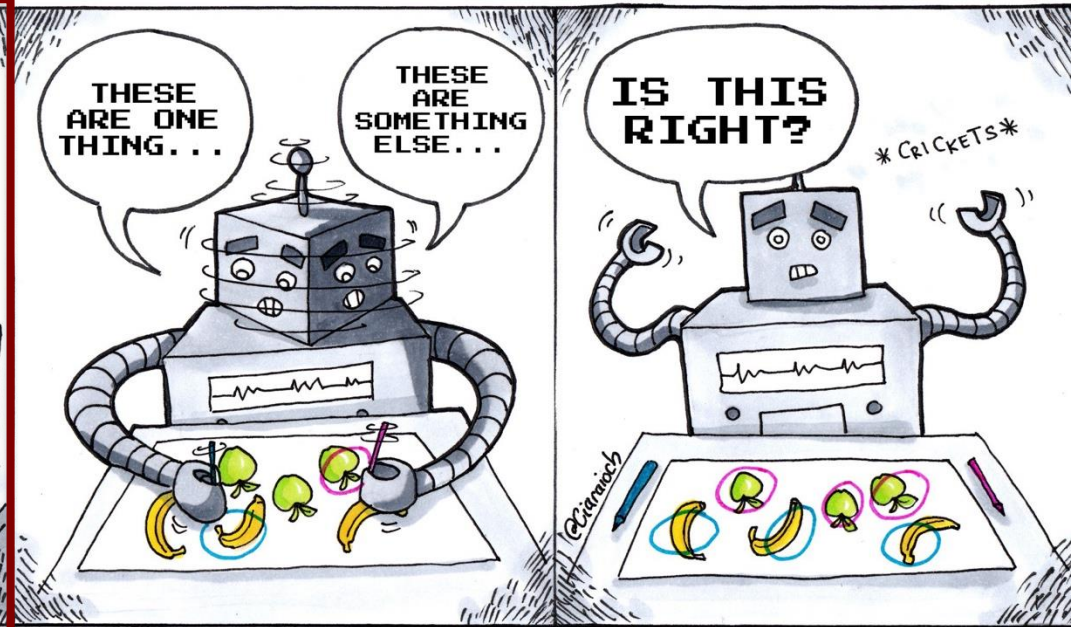


The teacher does not influence how the child learns the task but **observes the conclusions** the child can draw from solving it

Learning process: supervised vs unsupervised

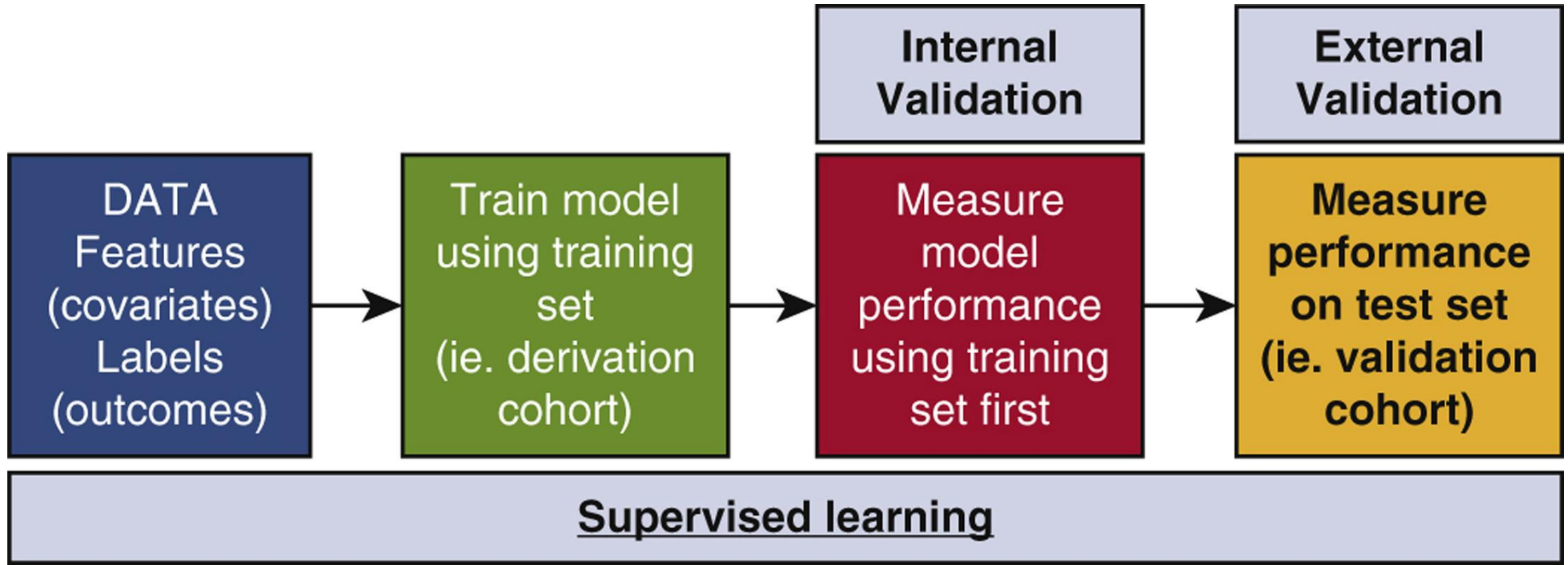


Supervised Learning



Unsupervised Learning

Supervised ML



1. STUDY POPULATION

DISCOVERY COHORT



Children with CeD
with nonspecific
clinical features



2014 – 2019

Sapienza University
of Rome, Italy

CONTROLS



Comparable for
age and sex
(frequency matching)



No individual
matching



Demographic,
anthropometric,
clinical and laboratory
data were collected
for all patients.



INCLUSION CRITERIA

-  ≤ 18 years old
-  No previous diagnosis of CeD
-  CeD patients with complete data for all features

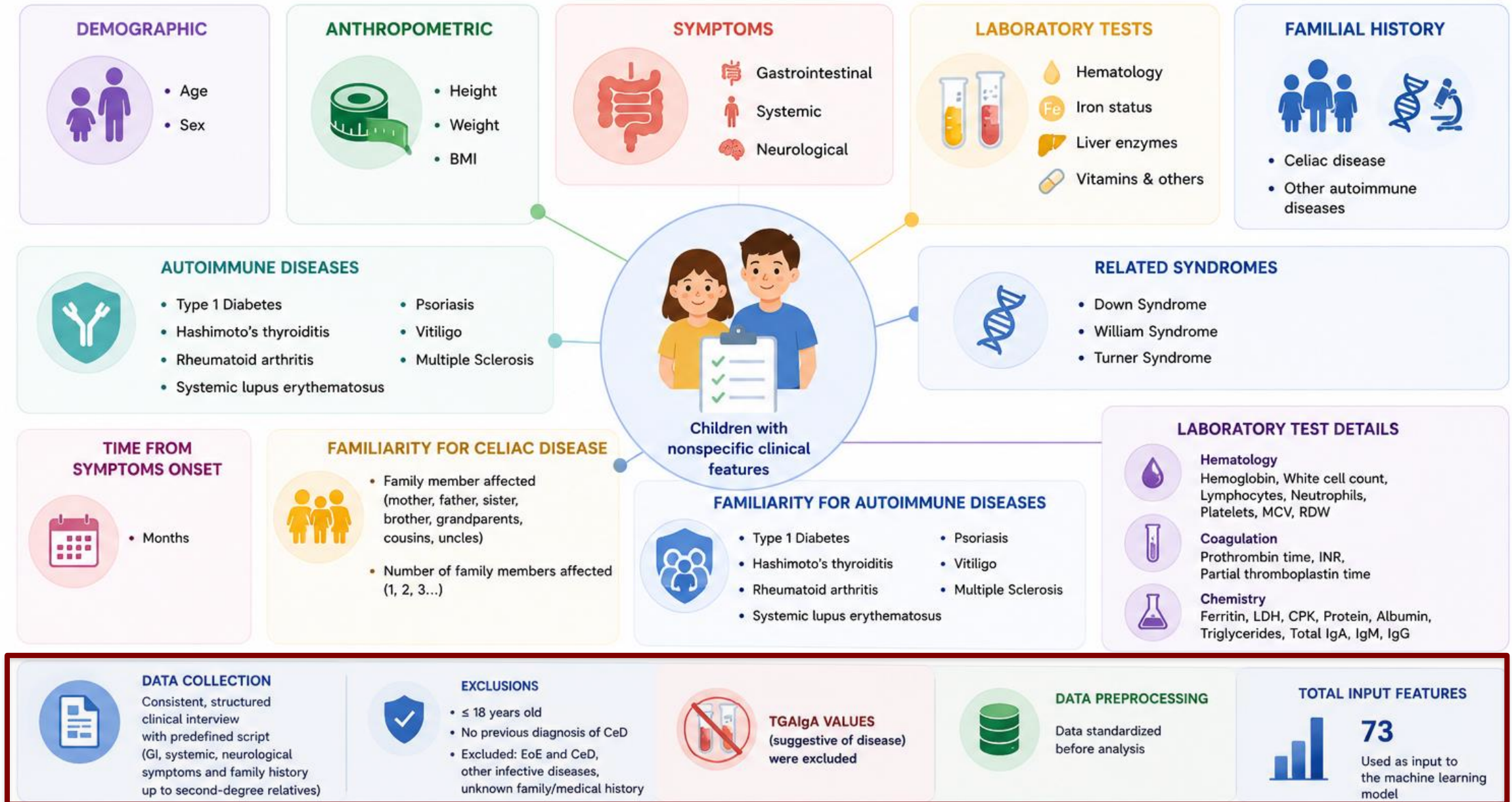


EXCLUSION CRITERIA

-  EoE and CeD
-  Other infective diseases
-  Unknown family and/or medical history

CLINICAL AND LABORATORY FEATURES CONSIDERED

Collecting the right information to predict celiac disease in children

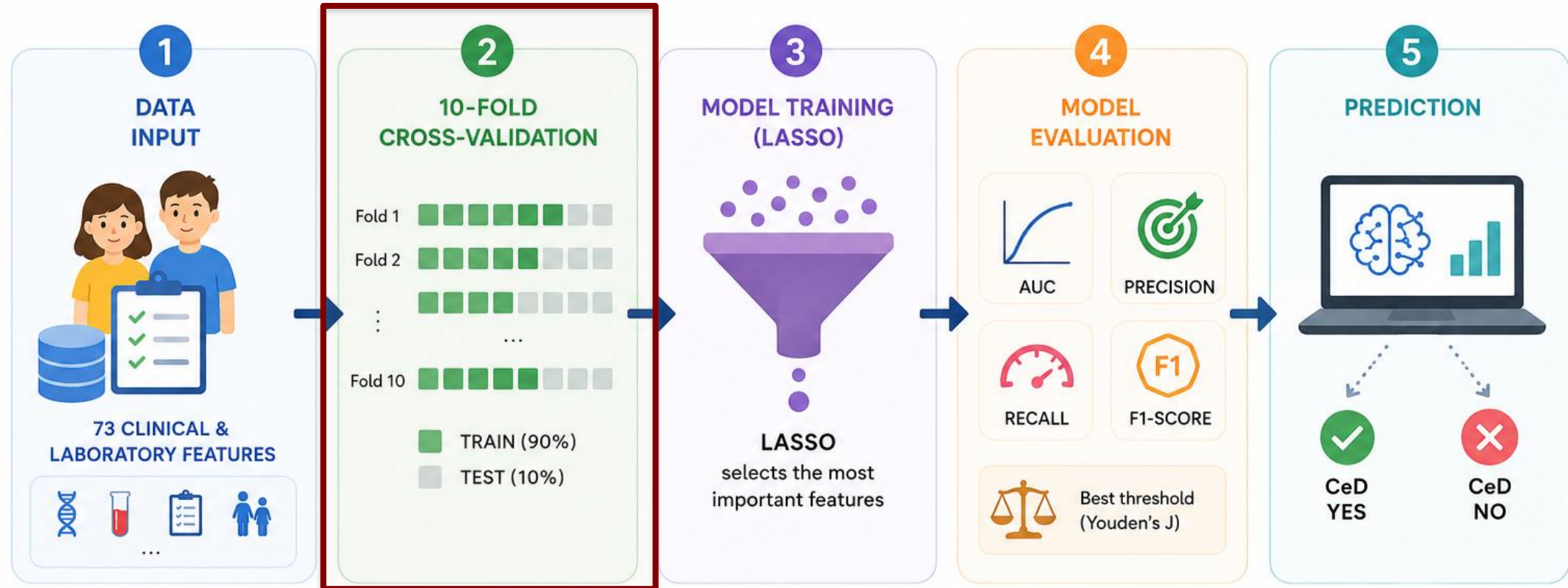


Learning process: how it works

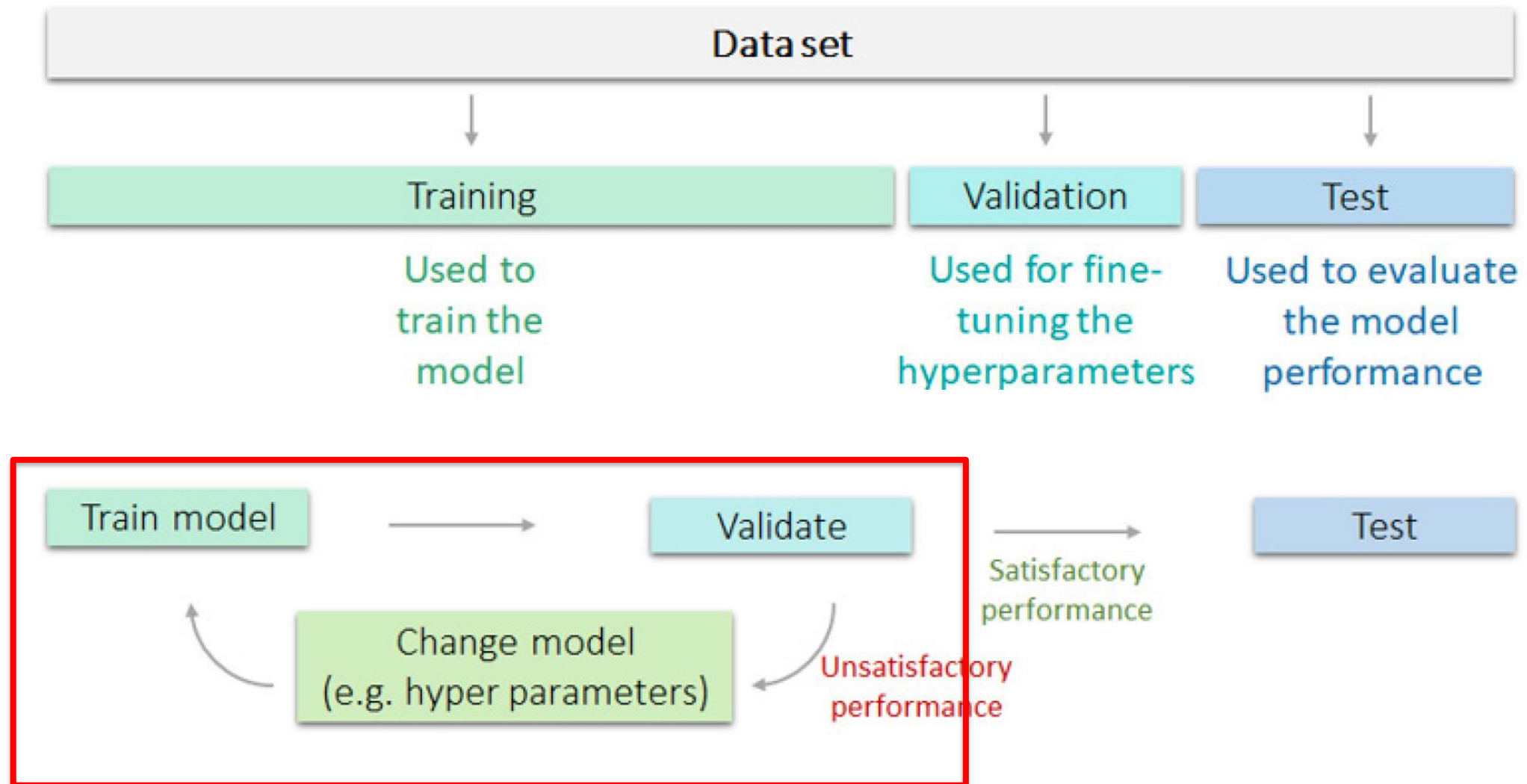


MACHINE LEARNING PIPELINE

Predicting Celiac Disease in Children



ML development: Training – Validation – Test

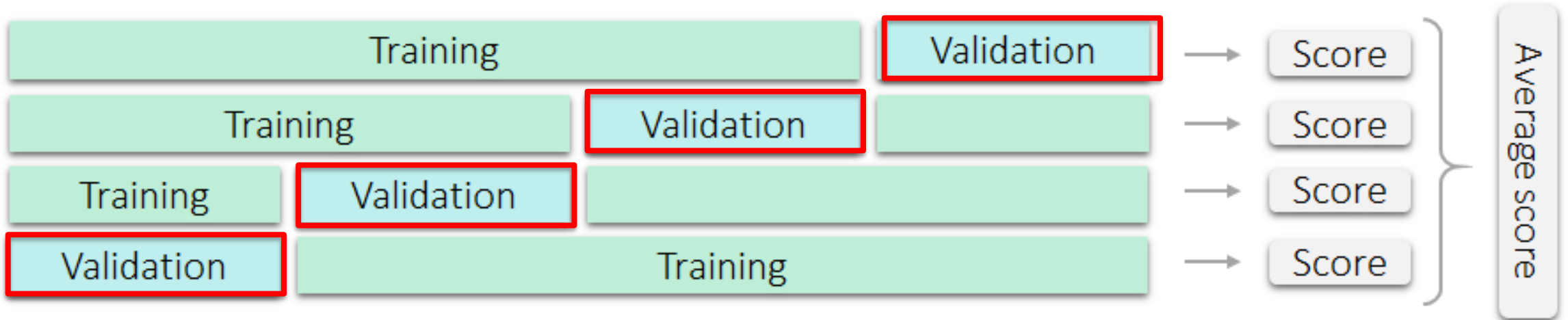


SHOW ME THE



DATA

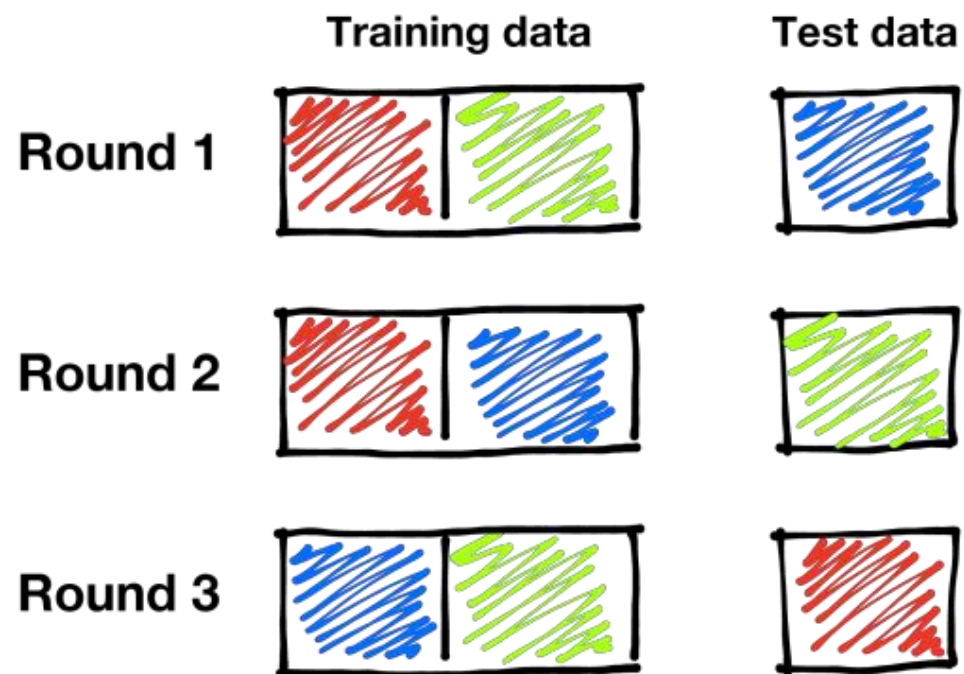
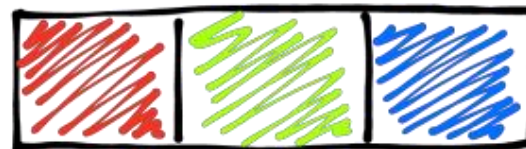
ML development: Cross-Validation



In K-fold cross-validation, the data used for training and validation are **split into K parts**, after which the model is subsequently trained with K-1 folds of **data and validated on the left-out fold**. This is repeated for all folds after which the scores are pooled.

ML development: Cross-Validation

Original data, divided into k parts



ML development: Cross-Validation

When you realize k-Fold Cross Validation can only validate your hyperparameters, not yourself..



HOW WE COMPARE MACHINE LEARNING MODELS

We evaluate each model across multiple complementary aspects to identify the best performer and ensure a fair comparison.

EVALUATION METRICS

1. ROC AUC

(Discrimination)



Ability of the model to distinguish between CeD patients and controls across all thresholds.

Higher is better ↑

2. F1 SCORE

(Balance)



Harmonic mean of precision and recall. Balances false positives and false negatives.

Higher is better ↑

3. ACCURACY

(Overall Correctness)



Proportion of total correct predictions (both CeD and controls).

Higher is better ↑

4. PRECISION (PPV)

(Positive Predictive Value)



Among predicted CeD patients, the proportion that are truly CeD positive.

Higher is better ↑

5. RECALL (SENSITIVITY)

(True Positive Rate)



Among actual CeD patients, the proportion that are correctly identified.

Higher is better ↑

MODELS COMPARED



LASSO

(L1 Regularization)



RIDGE

(L2 Regularization)



SVM

(Linear Kernel)



RANDOM FOREST

(Ensemble)



ADABOOST

(Boosting)



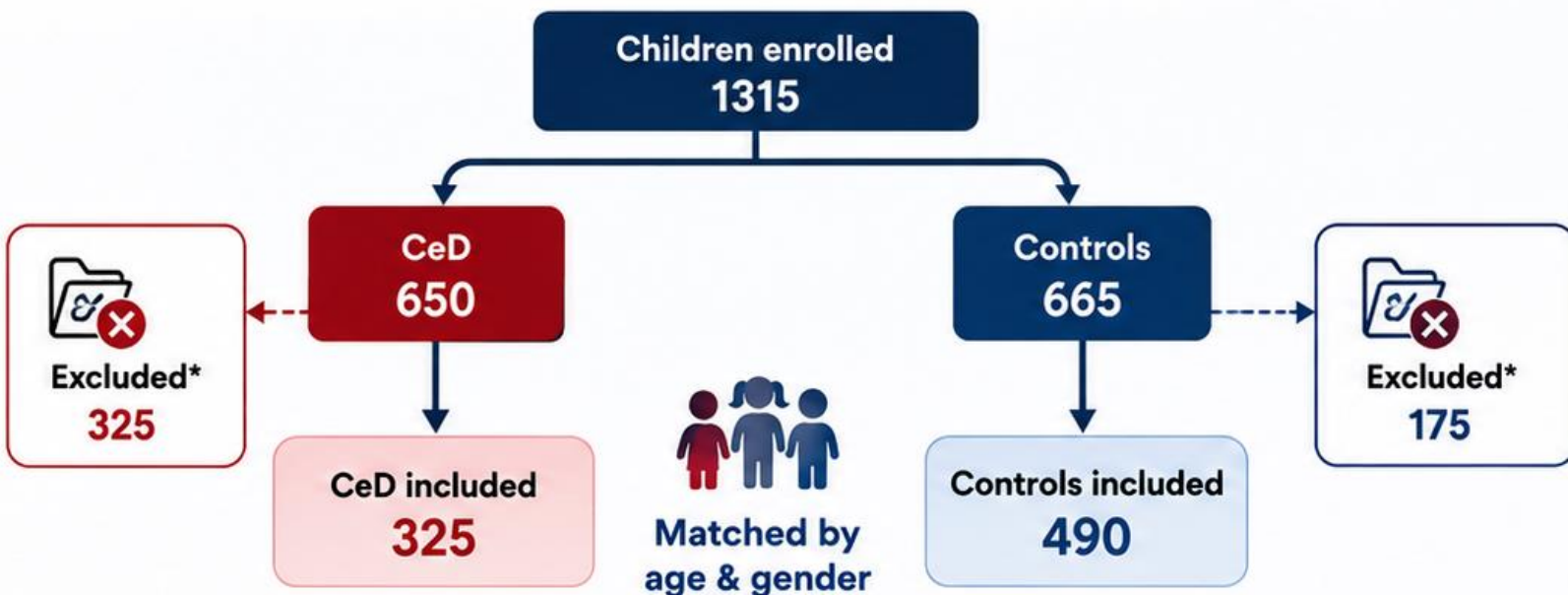
OUR GOAL

Identify the model that achieves the best predictive performance while providing reliable and clinically meaningful results.



3. STUDY DESIGN & COHORT

A discovery cohort of children with non-specific clinical features compared with **age- and gender-matched** controls



* 325 CeD and 175 Controls were excluded because they did not fill enrolled criteria listed in Table 1

FINAL ANALYTICAL COHORT

 **325** CeD patients **vs**  **490** Controls

KEY FEATURES

-  **73 non-serological clinical features** (symptoms, labs, family history, auxology)
-  **No anti-transglutaminase IgA values** included in the models
-  **Supervised machine learning models** with 10-fold cross-validation

DATA DOMAINS CONSIDERED (73 FEATURES)



Demographics & Auxology



Symptoms (Typical & Atypical)



Family history (CeD & Autoimmune)



Laboratory Tests

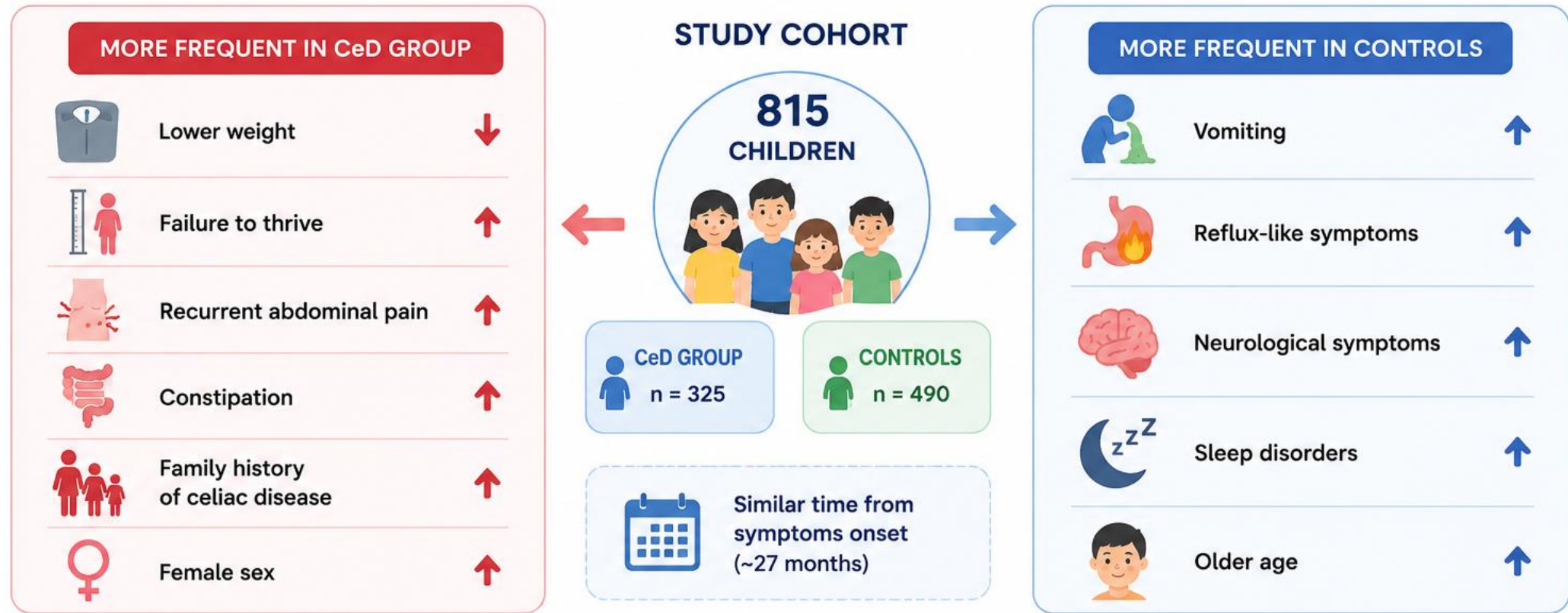


Associated Autoimmune Diseases



Time from Symptoms Onset

CeD vs CONTROLS: KEY DIFFERENCES AT A GLANCE



LABORATORY DIFFERENCES (CeD vs CONTROLS)



↓ Lower MCV
(Mean Cell Volume)



↓ Lower IgG



↓ Lower Prothrombin Time

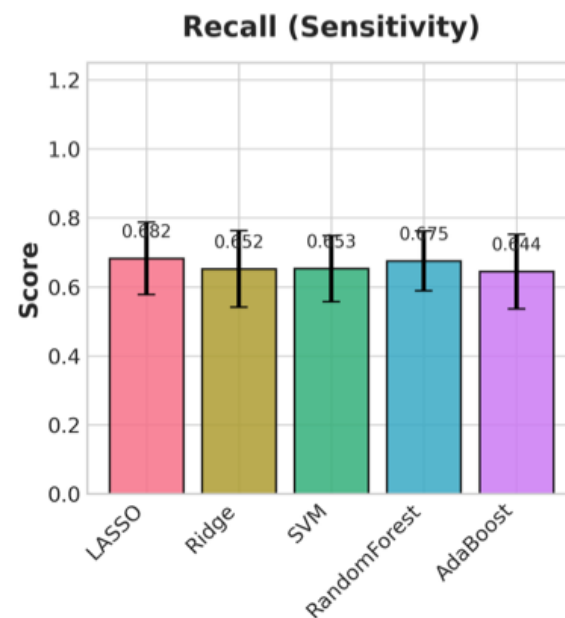
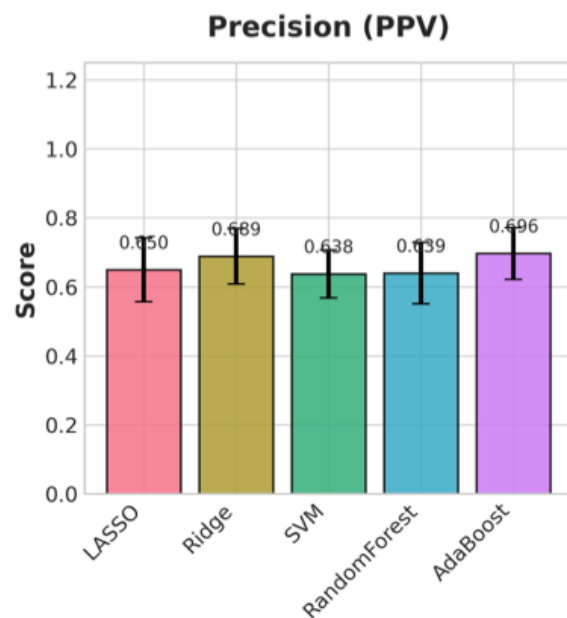
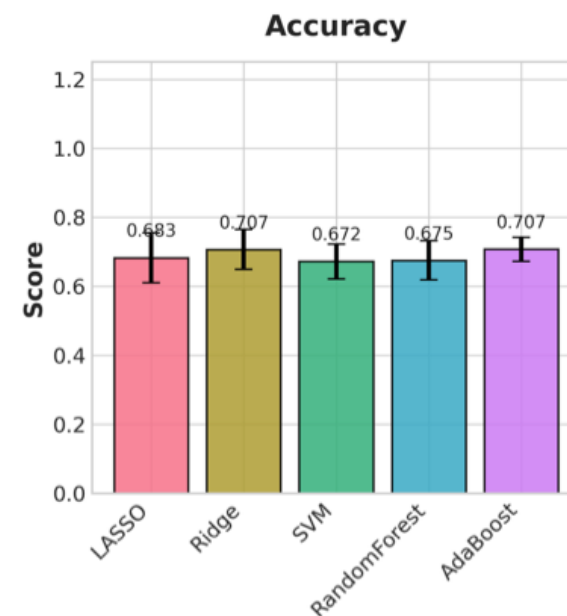
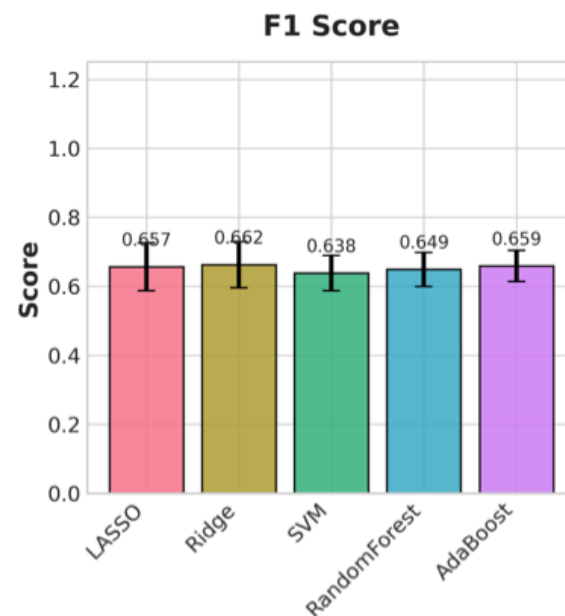
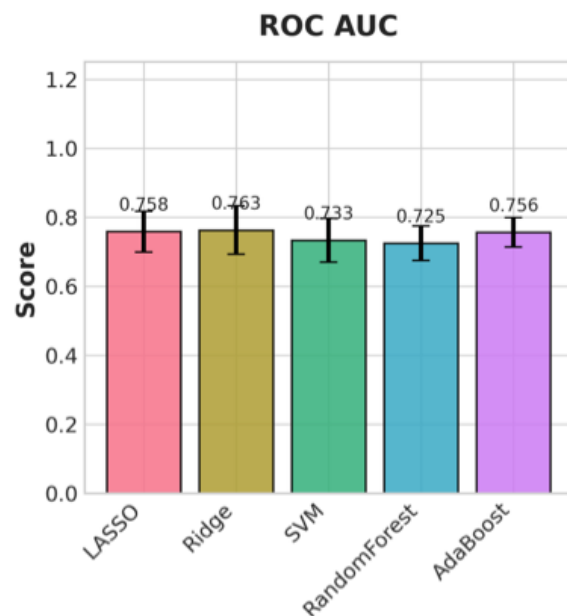


MACHINE LEARNING

identified patterns beyond classical celiac symptoms, improving the detection of celiac disease in children.



Machine Learning Model Performance Comparison



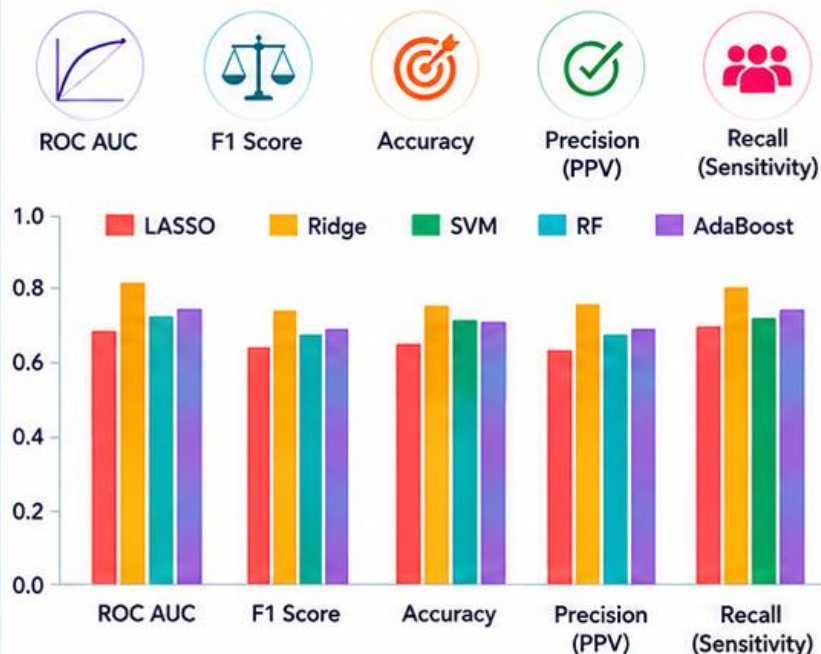
the Ridge classifier emerged as **the most balanced**, achieving the highest mean F1-score (0.662) and a mean AUC of 0.763.

LASSO: PREDICTION + INTERPRETABILITY

Slightly lower predictive metrics, but *unique feature selection* capability

1. MODEL COMPARISON

Performance across 5 metrics



Best overall performance:

Ridge



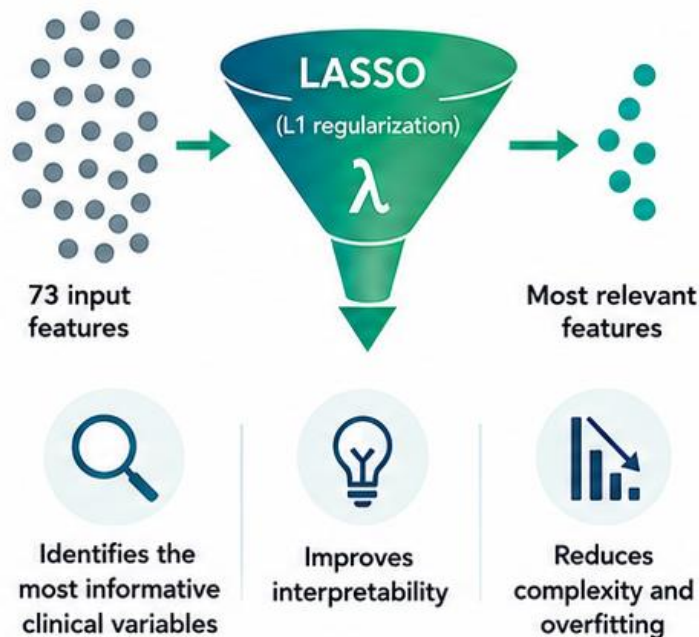
LASSO

Most stable recall

Ridge/AdaBoost achieve the highest predictive performance across most metrics.

2. WHY LASSO WAS CHOSEN

The only model with built-in feature selection (L1 penalty)



Enables discovery of key disease drivers beyond prediction alone.

3. RESULTS OF FEATURE SELECTION

Top 40 most important features (selected by LASSO)

Consistency (across folds)



FAMILY HISTORY (AUTOIMMUNE DISEASES)

- Hashimoto's thyroiditis
- Psoriasis



KEY SYMPTOMS

- Muscle pain
- Fatigue
- GERD-like symptoms



LABORATORY DEVIATIONS

- Hemoglobin (↓)
- Ferritin (↓)
- MCV (↓)
- IgG (↓)



High stability across cross-validation folds (Figure S2)



Clinical insights into CeD diagnostic drivers



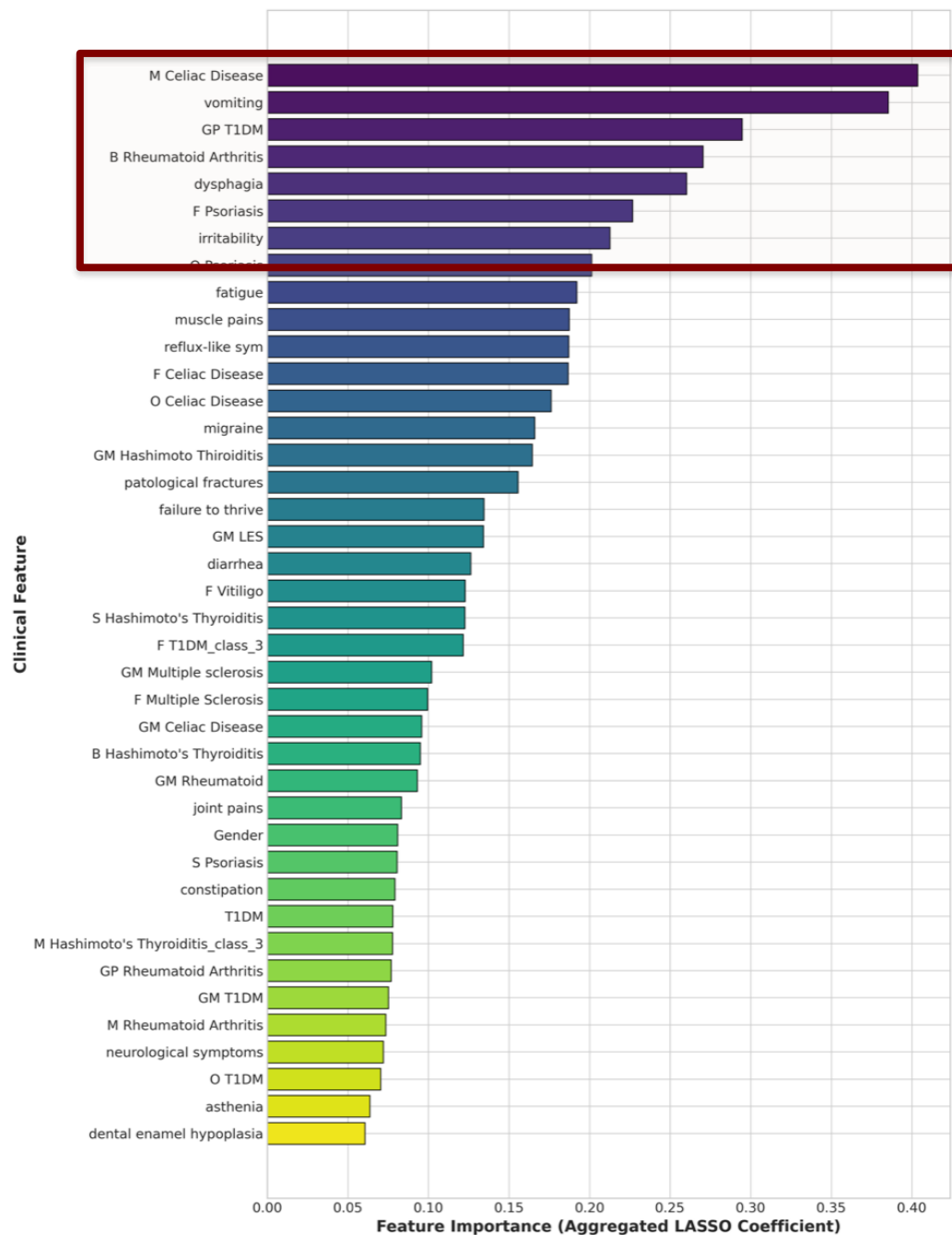
LASSO balances performance and interpretability, providing clinically meaningful insights.



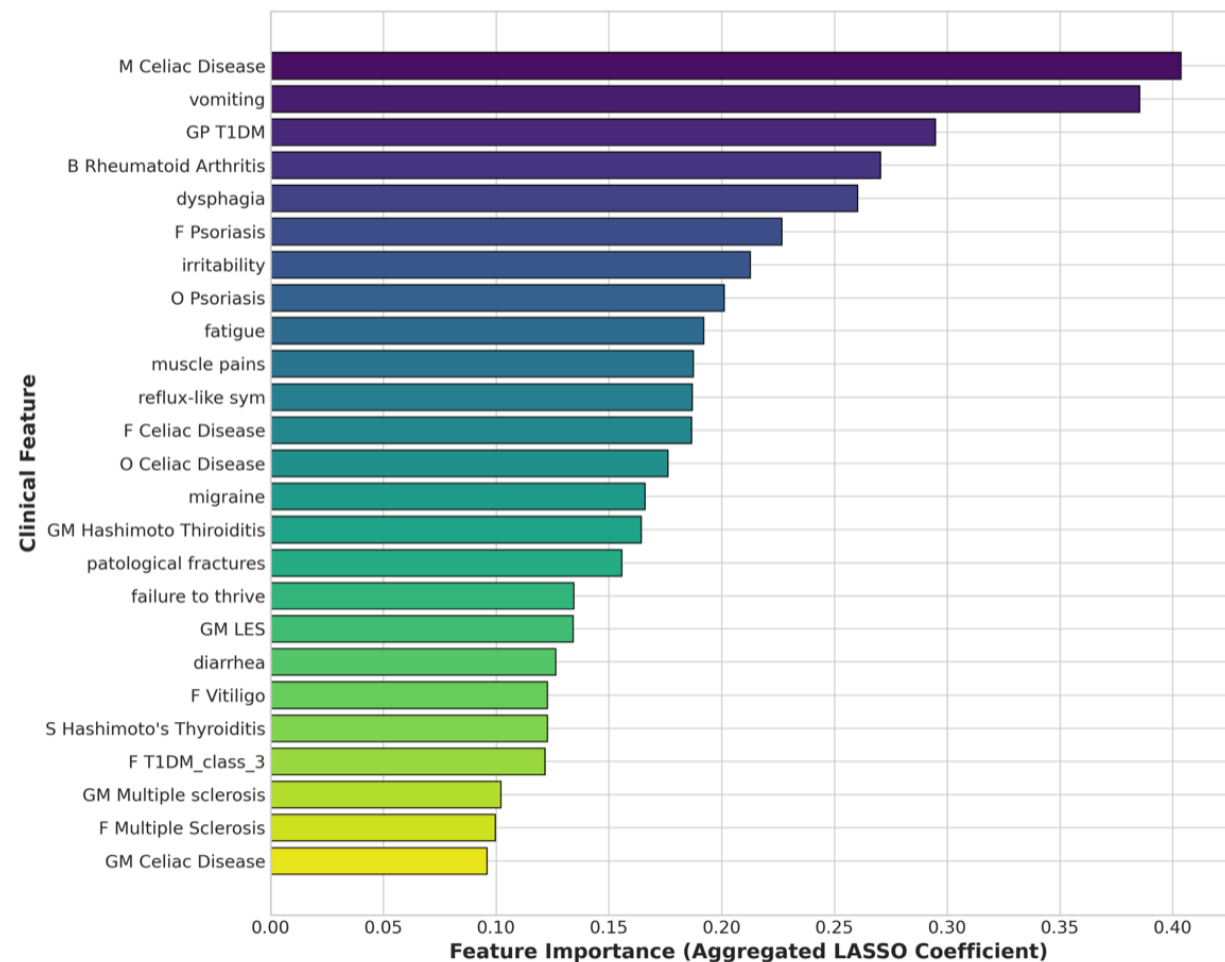
Prediction power + Understanding disease mechanisms
= **Better clinical decision-making**



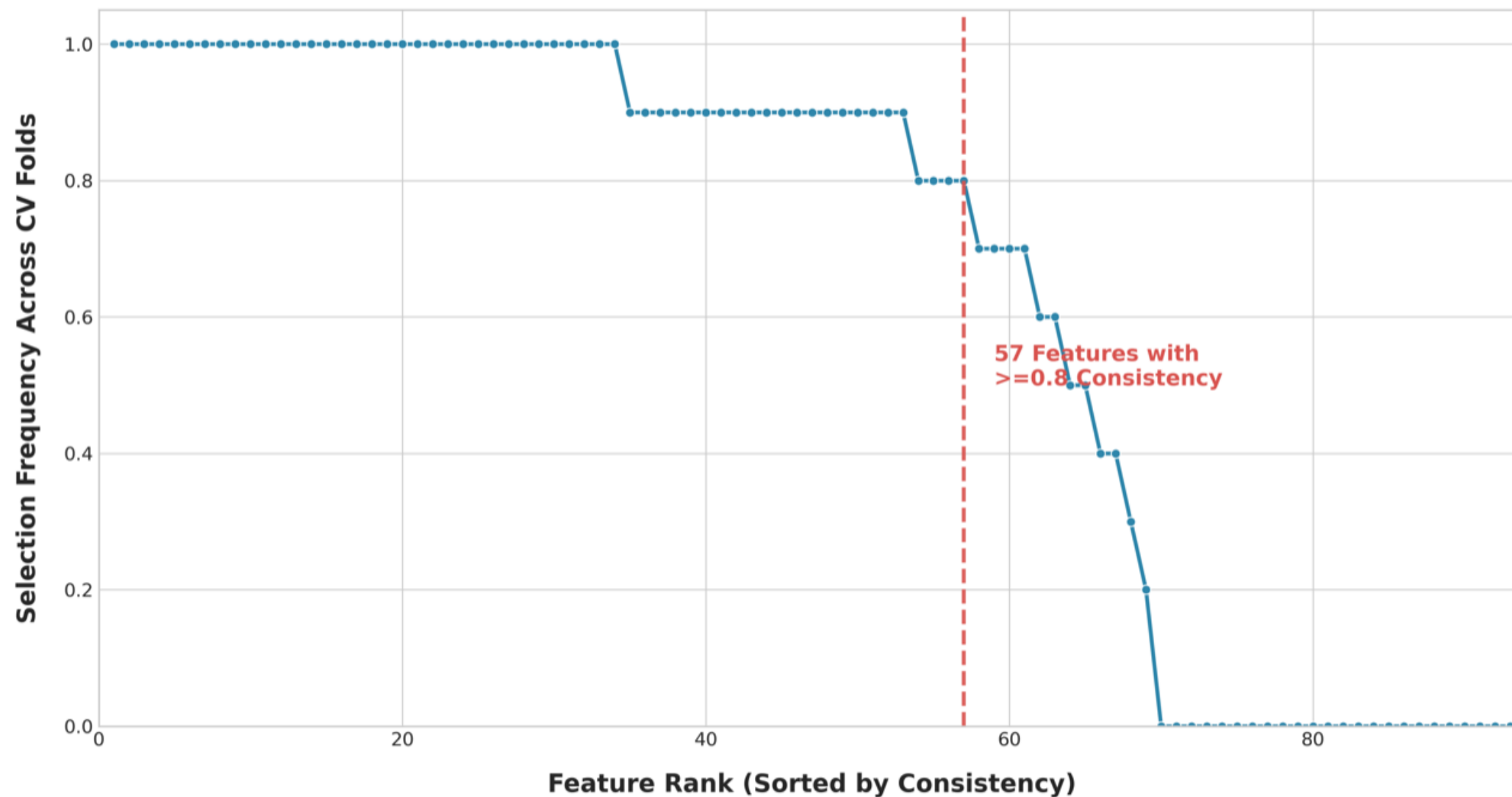
Top 40 Most Important Clinical Features



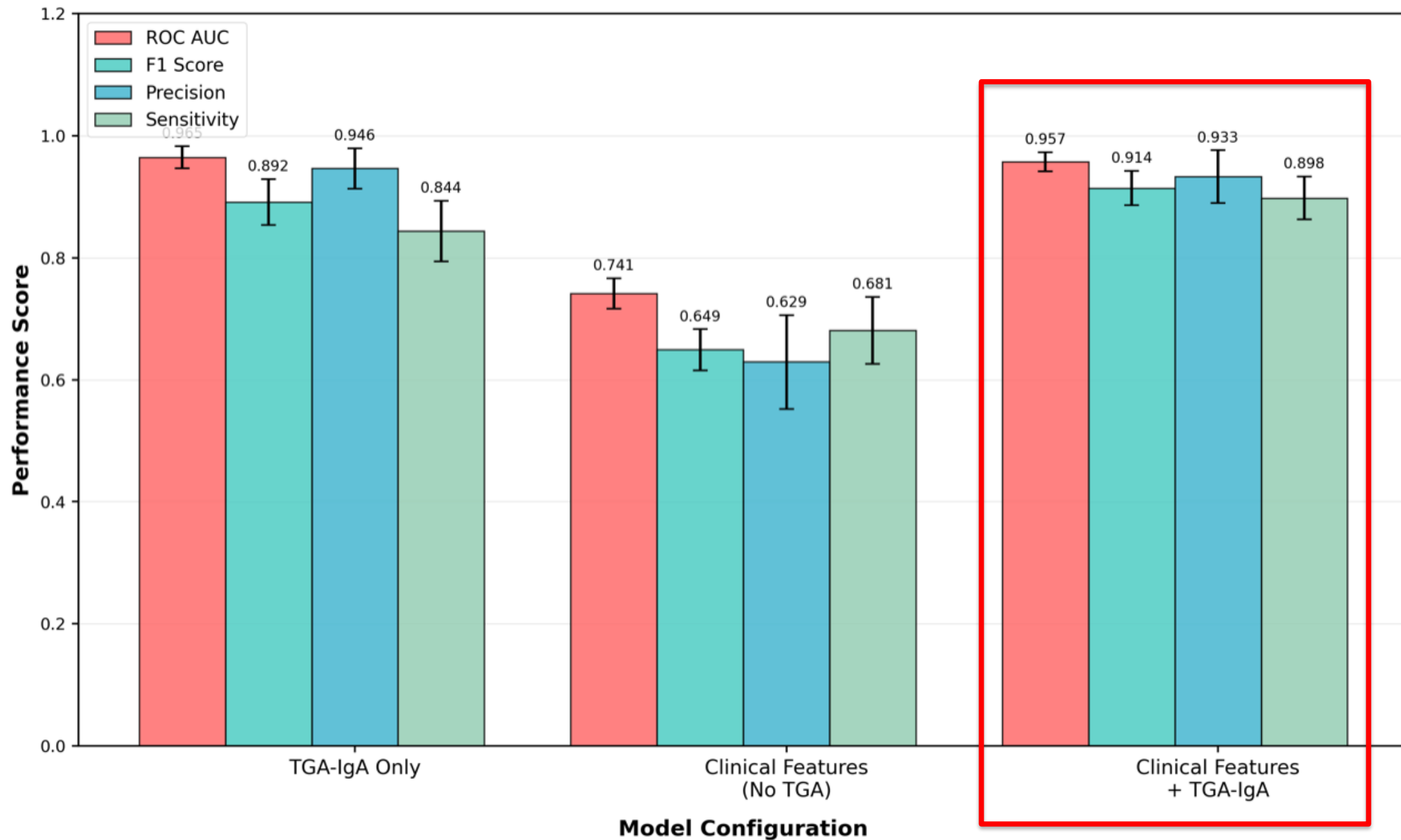
Top 25 Most Important Clinical Features



Consistency Decay of LASSO Features



TGA-IgA Benchmark Comparison



DISCUSSION

Finding the hidden celiac iceberg



WHAT WE FOUND

Machine Learning identified
40 stable non-serological predictors



Family history of autoimmune diseases

- Hashimoto's thyroiditis
- Psoriasis



Non-obvious symptoms

- Muscle pain
- Fatigue
- GERD-like symptoms



Routine laboratory abnormalities

- Ferritin (↓)
- Hemoglobin (↓)
- MCV (↓)
- IgG (↓)



Potential new **"red flags"** for screening



CLINICAL IMPACT



Child with vague symptoms

Nonspecific and non-classical features



ML-based risk assessment

Integrates 73 clinical, laboratory and family history features



Targeted TGA-IgA testing

Screen only children at higher risk



Earlier diagnosis

Timely treatment and better outcomes

Case-finding instead of mass screening



LIMITATIONS



Single-center cohort

Data from one referral center



Pediatric Italian population only

May limit generalizability to other populations



Clinical data depend on physician reporting

Subject to recall and reporting bias



Different laboratory platforms

Potential variability in laboratory measurements



No external validation cohort

Findings need replication in independent populations

Diagnosed CeD

Undiagnosed CeD



ML-assisted case finding

Discover the hidden patients



“

Machine Learning may help pediatricians recognize children who deserve celiac screening **before serology is ordered.**

”

2015



**SELF-DRIVING
CARS IN 2 YEARS**

CausalPython.io

2016

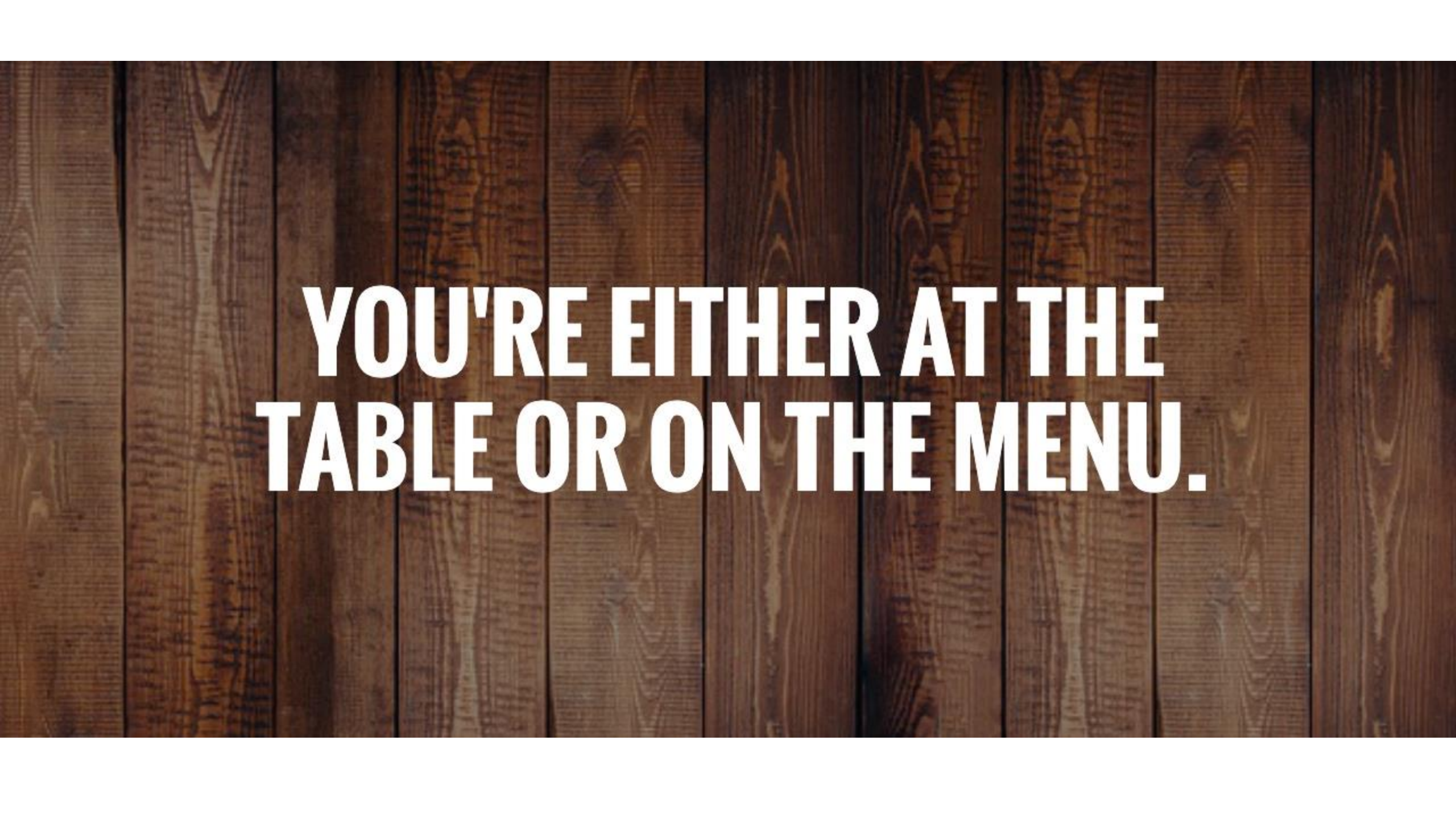


**RADIOLOGISTS
OBSOLETE
IN 5 YEARS**

2024



**RADIOLOGISTS
IN REGULAR CARS
DRIVING TO WORK**



**YOU'RE EITHER AT THE
TABLE OR ON THE MENU.**

Grazie



gastropediatria@uniroma1.it



UMBERTO I
POLICLINICO DI ROMA



SAPIENZA
UNIVERSITÀ DI ROMA