Prediction of ICU admission for COVID-19 patients: a Machine Learning approach based on Complete Blood Count data

Lorenzo Famiglini

Giorgio Bini

Anna Carobene

Andrea Campagner

Federico Cabitza

Introduction

Background on the topic

The **COVID-19** pandemic has had a large impact on public health systems around the world. For this reason there has been increasing interest for the **development of AI systems** for the disease management (**diagnosis + prognosis**).

Goals & Objectives

Most works have focused on the development of diagnostic systems, while the development of **prognostic models has been more scarce and largely centered on the death prediction task**: other important task exist, such as **ICU admission prediction**.

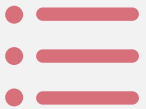
The goal of this work was the development of a ML-based tool to support clinicians in the prediction of ICU admission, so to help in the management of ICU allocation.

Dataset

Data collection

Emergency Department for COVID-19 of the San Raffaele Hospital (OSR), **Milan** (Italy). The data collection was performed **between February 19, and May 31**, 2020.

22 variables



Statistics

The **average age** of the patients was **63.5 ± 0.85** and the distribution of **biological sex** was **70.8%** males.

Class Imbalance

Skewed distribution in favor of the negative class: **27%** of the total observations.

SMOTE

4995 observations

For each instance (that is, **one day** of hospital stay for each given patient), the target corresponds to whether the given **patient would be admitted to the ICU within the next 5 days**.

TABLE I

COMPLETE LIST OF PREDICTIVE COVARIATES USED FOR THE MODEL DEVELOPMENT

Feature	**Unit of Measure**	**Missing rate (%)**
Sex	Male/Female	0
Age	Years	0
White Blood Cells (WBC)	$10^9/L$	0.4
Red Blood Cells (RBC)	$10^12/L$	0.4
Hemoglobin (HGB)	g/dL	0.4
Hematocrit (HCT)	%	0.4
Mean Corpuscular Volume (MCV)	fL	0.4
Mean Corpuscular Hemoglobin (MCH)	pg/Cell	0.4
Mean Corpuscular Hemoglobin Concentration (MCHC)	g Hb/dL	0.4
Erythrocyte Distribution Width (RDW)	CV%	0.5
Platelets (PLT)	$10^9/L$	0.4
Mean Platelet Volume (MPV)	fL	3.5
Neutrophils Count (NE—NET)	% — $10^9/L$	8.4
Lymphocytes Count (LY—LYT)	% — $10^9/L$	8.4
Basophils Count (BA—BAT)	% — $10^9/L$	8.4
Eosinophils Count (EO—EOT)	% — $10^9/L$	8.4
Monocytes Count (MO—MOT)	% — $10^9/L$	8.4

Methods

train/test split

80%-20% data split with the additional constraint that all observations pertaining to each given patient were all in the same data fold.

3 models

Two interpretable models, i.e., a **decision tree** and a (regularized) **logistic regression**; and a black-box **ensemble model**.

Experimental setup

Hyper-parameter selection, model training and validation were performed on the training set through a **10-time repeated 7-fold Cross-Validation.**

Hyper-parameter selection method

Sequential Model-Based Optimization (**SMBO**) approach.

$$F_2 = 5\frac{PPV \cdot Sensitivity}{4 \cdot PPV + Sensitivity}$$

Results

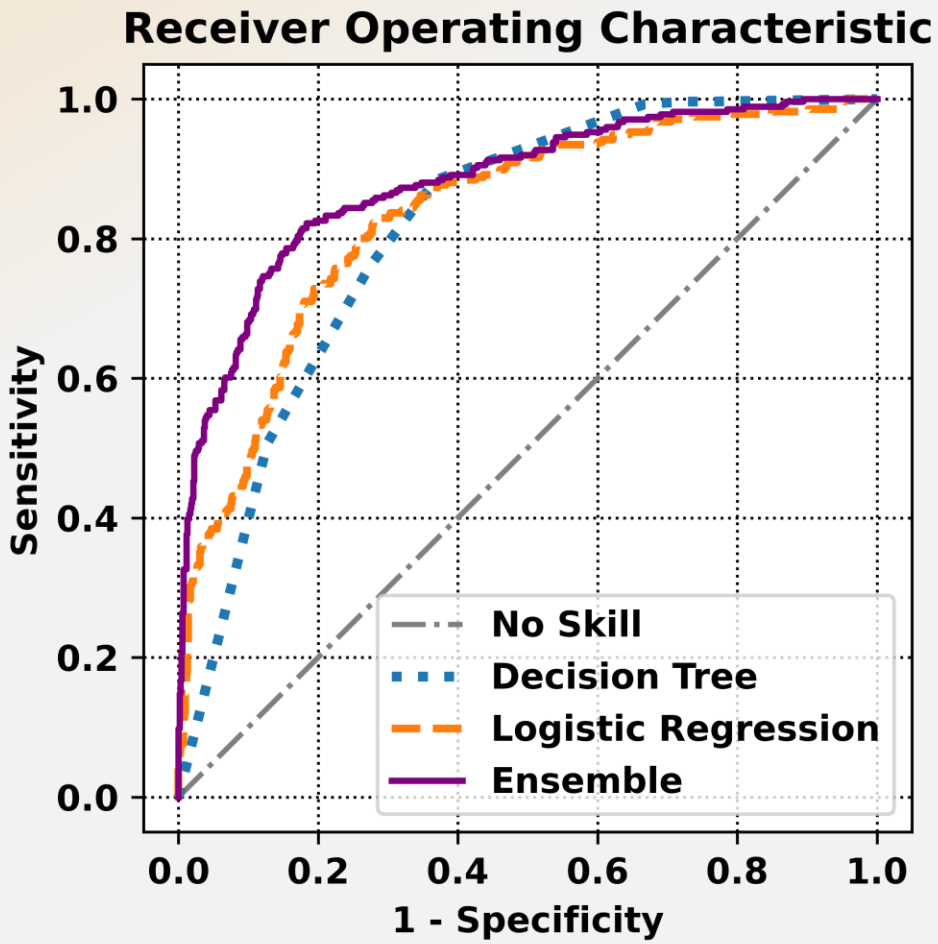


Fig.1 ROC Curve evaluated on the Test Set.

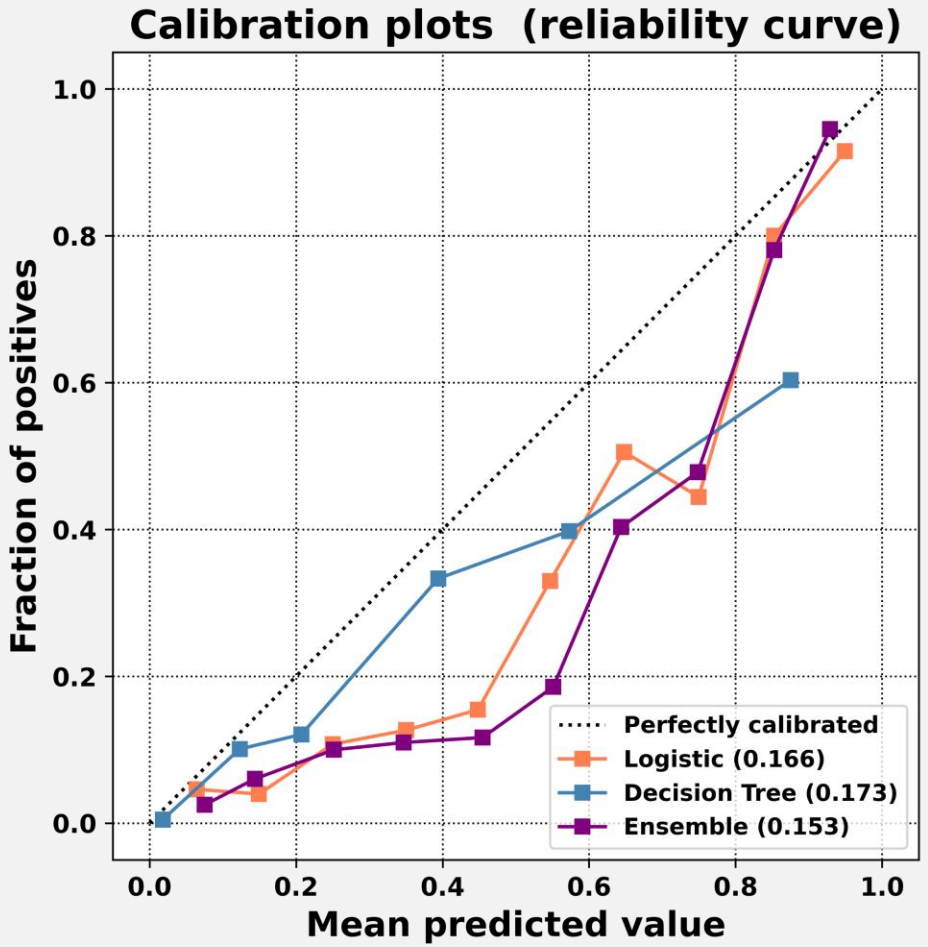


Fig.2 Reliability curve for each model.

Results

TABLE II
RESULTS OBTAINED ON THE TEST SET.

Model	Sensitivity	Specificity	AUC	F_2	Brier score	HC-AUC	HC-Sensitivity	HC-Specificity	HC-F_2	Coverage
Decision Tree	0.76	0.73	0.81	0.69	0.17	0.86	0.60	0.93	0.63	0.72
Logistic Regression	0.83	0.70	0.83	0.74	0.17	0.92	0.76	0.94	0.78	0.43
Ensemble	0.85	0.74	0.88	0.77	0.15	0.93	0.75	0.94	0.78	0.58

Models interpretation

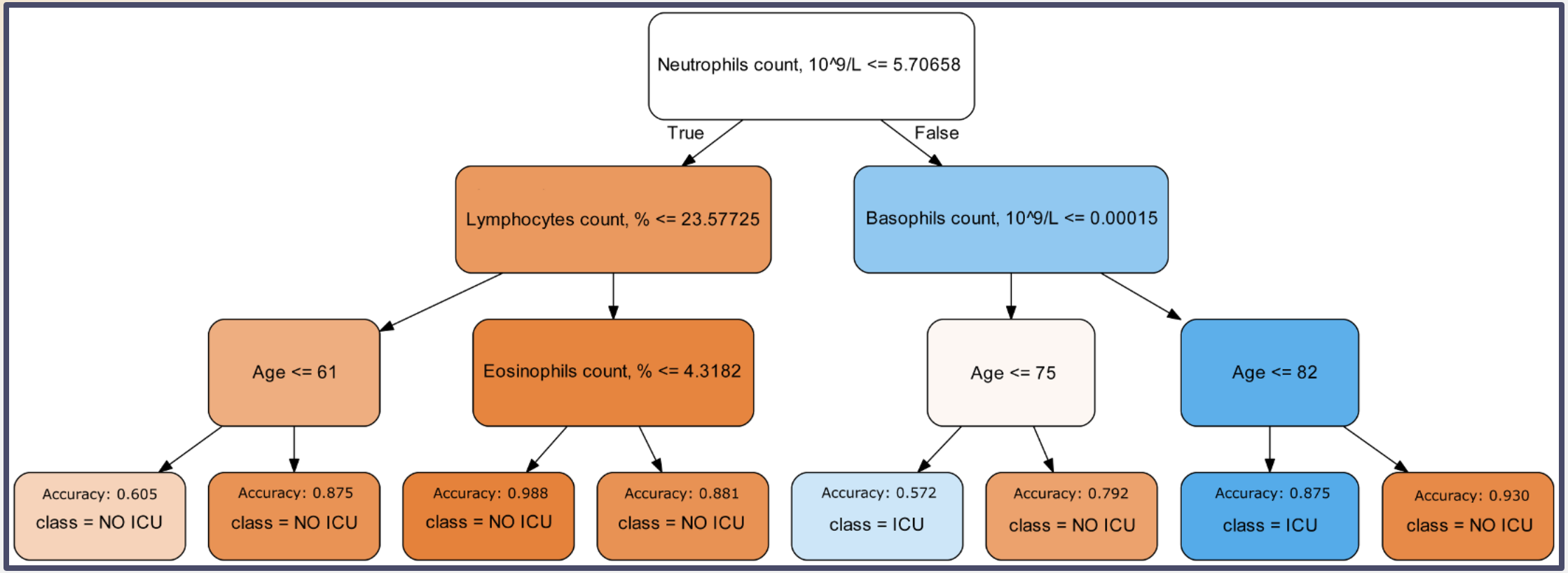


Fig.3 A graphical representation of the Decision Tree

Models interpretation

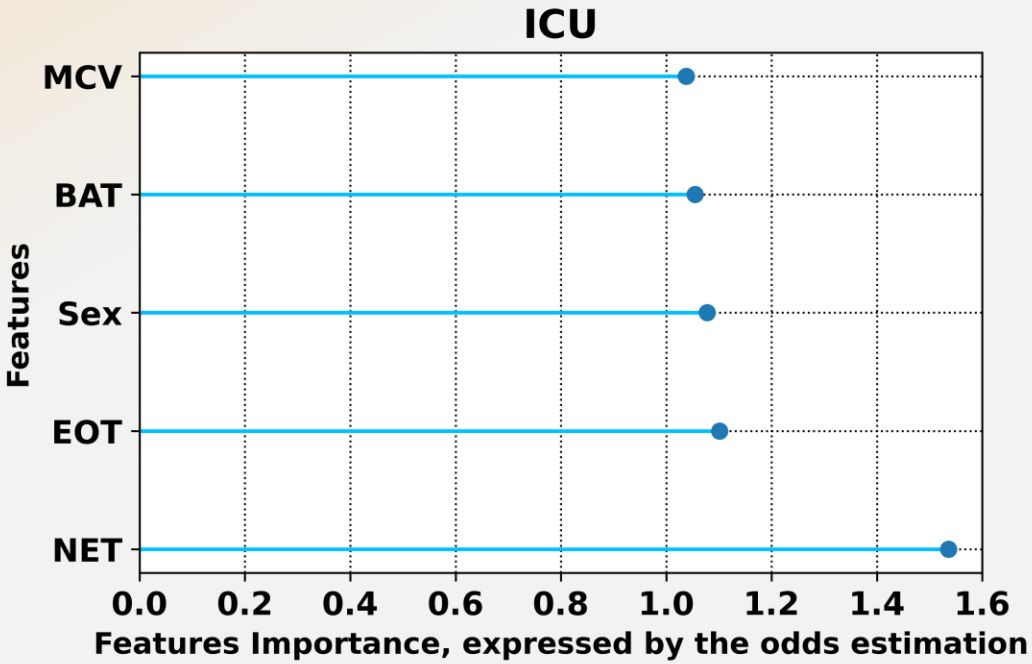


Fig.4 Feature importance based on Logistic Regression coefficients, for the positive class (that is, admission to ICU).

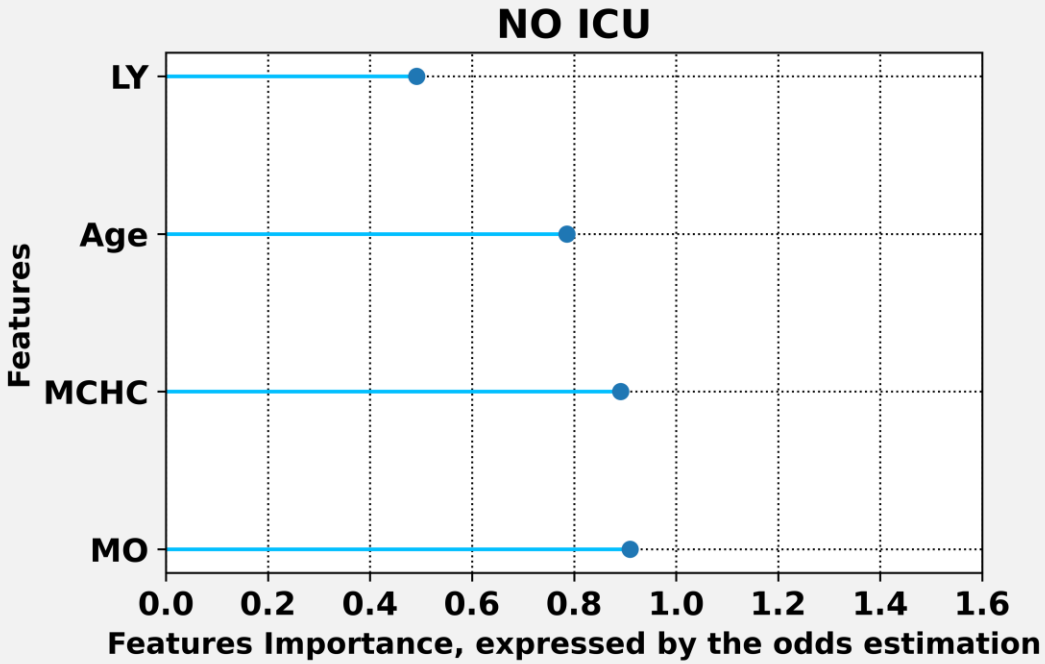


Fig.5 Feature importance based on Logistic Regression coefficients, for the negative class.

Conclusions

The proposed approach reported **good results**.

Our methods are **parsimonious**; for this reason, they can be **useful in resource-limited settings**, such as healthcare facilities.

Further Work

We aim to **externally validate** our models with data coming from **other hospitals and other time periods**.

THANK YOU !

a.campagner@campus.unimib.it