Cost-sensitive ordinal classification methods to predict SARS-CoV-2 pneumonia severity

Fernando García-García[1,*], Dae-Jin Lee[2,1], Pedro Pablo España Yandiola[3], Isabel Urrutia Landa[3], Joaquín Martínez-Minaya[4], Miren Hayet-Otero[5,6], Mónica Nieves Ermecheo[7], José María Quintana[8], Rosario Menéndez[9], Antoni Torres[10], Rafael Zalacain Jorge[11], and the COVID-19 & Air Pollution Working Group[12]

[1]Basque Center for Applied Mathematics (BCAM), Bilbao – Basque Country, Spain
[2]IE University, Madrid – Madrid, Spain
[3]Respiratory Service at Galdakao-Usansolo University Hospital, Galdakao – Basque Country, Spain
[4]Universitat Politècnica de València, Valencia – Valencian Community, Spain
[5]University of the Basque Country (UPV/EHU), Leioa – Basque Country, Spain
[6]Tecnalia BRTA, Derio – Basque Country, Spain
[7]BioCruces Bizkaia Health Research Institute, Barakaldo – Basque Country, Spain
[8]Research Unit at Galdakao-Usansolo University Hospital, Galdakao – Basque Country, Spain
[9]Pneumology Dept. at La Fe University Hospital, Valencia – Valencian Community, Spain
[10]Pneumology Dept. at Hospital Clínic, Barcelona – Catalonia, Spain
[11]Pneumology Service at Cruces University Hospital, Barakaldo – Basque Country, Spain
[12]A full list of collaborators in the COVID-19 & Air Pollution Working Group, with their respective affiliations, can be found at the end of the manuscript [*] [†]
[*]*Corresponding author*, e-mail: fegarcia@bcamath.org

Abstract

Objective: To study the suitability of cost-sensitive ordinal artificial intelligence-machine learning (AI-ML) strategies in the prognosis of SARS-CoV-2 pneumonia severity.

Materials & methods: Observational, retrospective, longitudinal, cohort study in 4 hospitals in Spain. Information regarding demographic and clinical status was supplemented by socioeconomic data and air pollution exposures.

We proposed AI-ML algorithms for ordinal classification via ordinal decomposition and for cost-sensitive learning via resampling techniques. For performance-based model selection, we defined a custom score including per-class sensitivities and asymmetric misprognosis costs. 260 distinct AI-ML models were evaluated via 10 repetitions of 5×5 nested cross-validation with hyperparameter tuning. Model selection was followed by the calibration of predicted probabilities. Final overall performance was compared against five well-established clinical

[*]Article accepted in *'IEEE Journal of Biomedical and Health Informatics'*. Manuscript received June 30, 2022; revised March 10, 2023 & December 27, 2023; accepted February 2, 2024.

[†]This research is supported by the Spanish State Research Agency (AEI) under the project *S3M1P4R* (*PID2020-115882RB-I00*), as well as by the Basque Government (EJ-GV) under the grant *'Artificial Intelligence in BCAM' 2019/00432*, under the strategy *'Mathematical Modelling Applied to Health'*, and under the *BERC 2022–2025* programme, and also by the Spanish Ministry of Science and Innovation: *BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIN / AEI / 10.13039/501100011033.*

severity scores and against a ‘standard’ (non-cost sensitive, non-ordinal) AI-ML baseline. In our best model, we also evaluated its explainability with respect to each of the input variables.

Results: The study enrolled n=1548 patients: 712 experienced low, 238 medium, and 598 high clinical severity. d=131 variables were collected, becoming d'=148 features after categorical encoding. Model selection resulted in our best-performing AI-ML pipeline having: *a*) no imputation of missing data, *b*) no feature selection (i.e. using the full set of d' features), *c*) *‘Ordered Partitions’* ordinal decomposition, *d*) cost-based reimbalance, and *e*) a Histogram-based Gradient Boosting classifier. This best model (calibrated) obtained a median accuracy of 68.1% [67.3%, 68.8%] (95% confidence interval), a balanced accuracy of 57.0% [55.6%, 57.9%], and an overall area under the curve (AUC) 0.802 [0.795, 0.808]. In our dataset, it outperformed all five clinical severity scores and the ‘standard’ AI-ML baseline.

Discussion & conclusion: We conducted an exhaustive exploration of AI-ML methods designed for both ordinal and cost-sensitive classification, motivated by a real-world application domain (clinical severity prognosis) in which these topics arise naturally.

Our model with the best classification performance exploited successfully the ordering information of ground truth classes, coping with imbalance and asymmetric costs. However, these ordinal and cost-sensitive aspects are seldom explored in the literature.

Keywords — Artificial intelligence, COVID-19, cost-sensitive classification, ordinal classification, SARS-CoV-2 pneumonia, severity prediction.

1 Introduction

After the rapid spread of the SARS-CoV-2 coronavirus and its outbreak into the global COVID-19 pandemic, the medical informatics and artificial intelligence-machine learning (AI-ML) communities dedicated large efforts to support medical decision-making towards high-quality healthcare for COVID-19 [1–3]. These initiatives ranged from predicting the spread of the disease [4, 5] or identifying populations at risk [6, 7], to diagnostic [8, 9] and/or prognostic tools [10–13].

In this context, medical informatics may provide pulmonologists with assistance for personalized, effective and efficient, data-backed therapeutic guidance [14].

1.1 Motivation

A ‘living’ systematic literature review [1] identified 107 AI-ML models for COVID-19 disease prognosis: 39 focused on predicting mortality risk, whereas 28 were for clinical progression. Prognoses were often either dichotomous by nature (e.g. admission or not to intensive care units, ICU [15]), or dichotomized. However, our pulmonologists considered it relevant to assess deterioration by distinguishing multiple levels of severity. Such a clinical question motivated us to investigate tailored AI-ML strategies able to address –by design– two key aspects of the motivating task:

- Having three severities entails a natural intrinsic order in the ground truth classes. Thus, here we formalized mathematically our task as an ordinal classification problem (*aka* ordinal regression) [16].
- The various types of misprognoses imply different consequences for patient safety, a fact which called upon the use of cost-sensitive learning [17].

These two choices entail –by themselves, as well as combined– methodological novelty with respect to the ‘standard’ nominal approach (either binary or even multi-class but non-ordinal) adopted by the majority of AI-ML literature in the context of COVID-19 [1].

The data collected to characterize each patient’s case comprised mostly demographic (e.g. sex, age) [18] and clinical information (e.g. blood analytics) as the core explanatory variables for the AI-ML systems to learn patterns in prognosis. In addition, our Working Group decided to supplement such information with extra factors, which had been consistently reported in the literature to aggravate health outcomes (in general, but also for COVID-19 in particular): socioeconomic inequalities [19, 20] and the exposure to outdoor air pollution – mainly particulate matter and NO_2 [21–25]. Population from deprived strata and/or regions have been reported to suffer from more severe consequences of COVID-19, and to be more prone to get infected [20, 26, 27], due to their living conditions, baseline health status, etc. Besides, air pollution may also predispose to chronic diseases (respiratory and cardiometabolic) [23] which can become aggravated by COVID-19: decreasing subjects’ immune response, facilitating viral entry and SARS-CoV-2 infection [24]. Chronic exposure to various pollutants (notably $PM_{2.5}$ particulate matter) was found to correlate with alveolar ACE-2 receptor over-expression, leading to more severe infections [25]. Furthermore, for the cohort studied here, our Working Group found statistically significant relations between: *a*) chronic exposure to pollutants (primarily nitrogen oxides NO_2, NO, NO_X, also PM_{10}); and *b*) SARS-CoV-2 pneumonia mortality, as well as with biomarkers of inflammation and gas exchange [28].

2 Materials & Methods

We adopted the *'IJMEDI checklist for assessment of medical AI'* [29] to guide the reporting of our study.

2.1 Data collection

We conducted an observational, retrospective, longitudinal, cohort study with a multi-center setup, in four major general hospitals from three different geographical territories in Spain: Catalonia (Clínic Hospital, in Barcelona), Valencian Community (La Fe Hospital, in Valencia), and the Basque Country (Galdakao-Usansolo and Cruces Hospitals, in Galdakao and Barakaldo). The study was approved by the corresponding Ethics Committees for Clinical Research (codes: HCB/2020/0273, 20-122-1, PI 2019090, PI 2020083), and it was carried out in adherence to the relevant guidelines and regulations. Only participants who voluntarily gave written informed consent were enrolled.

The inclusion criterion was adult patients ($\geq$18 years old) admitted to in-hospital stays due to SARS-CoV-2 pneumonia during the first wave of COVID-19 in Spain: between mid-February and end-May 2020. Requirements for COVID-19 pneumonia diagnosis were both a positive microbiological test (positive DNA amplification test by PCR for SARS-CoV-2), as well as compatible chest imaging findings (radiography, tomography).

A posteriori examination of the clinical records allowed us to determine the ground truth severity in their evolution. Our pulmonologists defined three target ordinal classes via systematic objective criteria. The HIGH severity group comprised patients who either: *a*) died intra-hospital or within 30 days after admission; or *b*) required major respiratory aids/aggressive treatments (high flow oxygen therapy, non-invasive mechanical ventilation, orotracheal intubation, extracorporeal membrane oxygenation, hemofilter, and/or vasoactives); or *c*) were admitted to ICU –including 'intermediate' respiratory ICUs–; or *d*) suffered major clinical complications (e.g. distress, shock). The MEDIUM severity group was formed by patients who either: *a*) stayed in the hospital for at least 14 days, or *b*) suffered intermediate complications (e.g. pulmonary embolism, congestive heart failure, neurological deterioration, etc.); whereas the LOW severity group comprised the rest of the patients, whose clinical evolution was thus favorable.

A broad set of variables were collected to characterize each patient's case (Section 3.1 for further details). These included: *a*) demographics (e.g. age, sex, body mass index, residence in a nursing home); *b*) comorbidities (i.e. pre-existing conditions); *c*) symptoms and physiological status during the preliminary emergency episode; and *d*) results from baseline examinations at the time of hospitalization (laboratory blood analytics, arterial gas tests, etc.).

Besides, we incorporated extra variables describing the socioeconomic situation in each patient's postcode of residence: average income, mean age, percentage of the population under 18 and over 65 years old, etc. These data were obtained from the most recent public census by the Spanish National Statistical Institute (INE, 2019) [30].

We also obtained daily air pollution measurements, throughout 2019 and during the relevant part of 2020, to characterize acute and chronic exposures to pollutants per geographical location. These data were published by the air quality agencies from the corresponding territorial authorities [31–34], for eight main pollutants: PM_{10}, $PM_{2.5}$, O_3, NO_2, NO, NO_X, SO_2, and CO.

2.2 Data preparation & pre-processing

As part of a preliminary curation stage to guarantee data quality and integrity, we disregarded *a priori* any variable suffering from $\geq$60% missing values (Figure 1). For the remaining variables, continuous measurements spanning several orders of magnitude (e.g. concentrations from the blood tests) were $\log_{10}$-transformed. Discrete categorical variables (without intrinsic order, e.g. type of bronchological comorbidity) were transformed into binary features via 'one-hot' encoding; whereas discrete ordinal variables (e.g. CURB-65 pneumonia score) were treated as integers.

Regarding socioeconomic information, since the original data were published per census districts, we re-interpolated them per postcode of residence (i.e. the available information). We first computed the portion of geographical area from each census polygon within a certain postcode polygon and then carried out a weighted sum of values.

For air quality, measurements were available only at the locations of surveillance stations. Thus, we estimated day-to-day pollution levels per postcode via Bayesian Generalized Additive Models (BGAMs) [35, 36], computing the distribution of exposures as a function of latitude, longitude, and elevation. In this study, we considered chronic exposure as the pollution throughout 2019; whereas acute exposure was across the 7 days before each patient's admission date.

Figure 1 depicts an overview of the structure of our dataset, along with a schematic of our compound AI-ML workflow, including the algorithmic choices for each step (Sections 2.3 to 2.10).

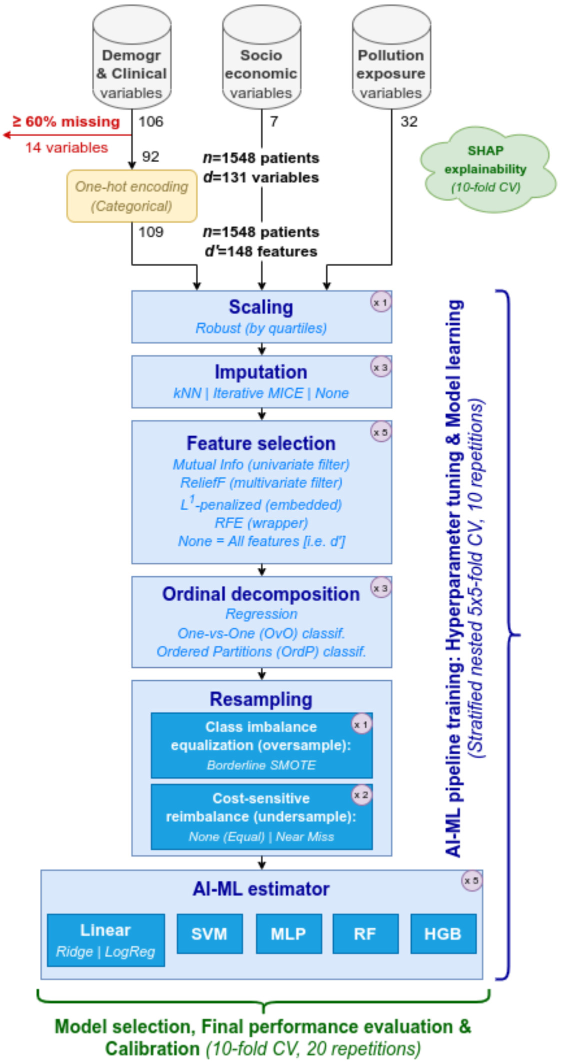

Demogr & Clinical variables
Socio economic variables
Pollution exposure variables
≥ 60% missing
14 variables
106
92
7
32
n=1548 patients
d=131 variables
One-hot encoding (Categorical)
109
n=1548 patients
d'=148 features
SHAP explainability (10-fold CV)
Scaling
Robust (by quartiles)
x 1
Imputation
kNN | Iterative MICE | None
x 3
Feature selection
Mutual Info (univariate filter)
ReliefF (multivariate filter)
L1-penalized (embedded)
RFE (wrapper)
None = All features [i.e. d']
x 5
Ordinal decomposition
Regression
One-vs-One (OvO) classif.
Ordered Partitions (OrdP) classif.
x 3
Resampling
Class imbalance equalization (oversample):
Borderline SMOTE
x 1
Cost-sensitive reimbalance (undersample):
None (Equal) | Near Miss
x 2
AI-ML estimator
x 5
Linear
Ridge | LogReg
SVM
MLP
RF
HGB
AI-ML pipeline training: Hyperparameter tuning & Model learning
(Stratified nested 5x5-fold CV, 10 repetitions)
Model selection, Final performance evaluation & Calibration (10-fold CV, 20 repetitions)

Figure 1: Workflow – Diagram of our methodology. In red, preliminary quality assurance discarded variables with ≥60% missing values. In golden, data pre-processing. In blue, steps forming the AI-ML pipelines. In green, additional procedures.

2.3 AI-ML pipelines

Methodologically, our proposal here consisted in examining compound AI-ML models (*aka* pipelines), whose stages (Figure 1) were specifically designed to tackle challenges posed by our SARS-CoV-2 pneumonia dataset, as well as by the nature of our cost-sensitive ordinal classification task.

After data pre-processing (Section 2.2), we performed input scaling. It aimed to guarantee that per-feature distributions have commensurate magnitudes: within comparable ranges of variation (e.g. regardless of their physical units of measurement), as commensurability is known to help in the convergence of many AI-ML algorithms. Instead of the archetypical *'standardization'* (i.e. subtract the sample mean, then divide by the sample standard deviation), here we opted for a so-called *'robust'* scaling: i.e. subtract the sample median (quartile q_2), then divide by the inter-quartile range ($q_3 - q_1$). The motivation for this choice is that quartiles are less sensitive to potential outliers than mean and deviation. We did not apply any further explicit outlier detection scheme.

Secondly, we proposed an imputation stage to cope with missing values. This is necessary because most AI-ML estimators (Section 2.6) are unable to handle missing data. We examined two different choices for an imputation strategy:

a) k-nearest neighbors (kNN), where an instance's missing values are estimated by the weighted average of the values for that same feature across the k instances closest to it ('neighbors') [37, 38]; and

b) Iterative (MICE), where the imputation formula for the missing values of a certain feature is computed as a regression function of the n_f other features most correlated to it [39, 40].

In addition, and given that one of the AI-ML estimators (HGB, see Section 2.6) is capable of learning with missing data, we also studied pipelines without any imputation. None of these choices altered the dimensionality of data: neither in terms of the amount of instances nor in the number of features.

As a third stage in our pipelines, we examined various feature selection techniques. The motivation was to ameliorate the *'curse of dimensionality'* issue [41], compacting the representation of the relevant information in a subspace of lower dimensionality. This may facilitate learning for the subsequent AI-ML estimators. In our complementary work [42], we conducted an exhaustive examination of 166 distinct feature selection strategies for this scenario, attending to objective and systematic criteria on selection performance (mostly, bootstrapped stability in the subset of selected features). Out of those 166 alternatives, here we employed the top four:

a) Mutual information (MI) univariate filter [41],

b) ReliefF multivariate filter [43],

c) L^1-penalized regularization for embedded selection [41],

d) Recursive feature elimination (RFE) wrappers [41].

From these, only ReliefF can handle missing data. Besides, given that some AI-ML estimators (e.g. tree bagging and boosting) may not always suffer from the *'curse of dimensionality'*, we also considered a case without feature selection, i.e. working with the full set of features.

As additional stages (Figure 1), we proposed: *a*) ordinal decomposition for ordinal classification (Section 2.4); *b*) instance resampling to cope with class imbalance, as well as for cost-sensitive learning (Section 2.5); and *c*) the AI-ML supervised estimators themselves (Section 2.6).

2.4 Ordinal classification

For the motivating scenario, our pulmonologists considered it relevant to discern Q=3 severity levels, which entail an obvious natural ordering. In the classical 'standard' AI-ML strategies, one would straightforwardly use nominal classification techniques, which are not designed to account for such an ordered structure of the ground truth classes. Thus, the underlying –and potentially enhancing– information about class order may remain unexploited.

Conversely, here we focused on solutions explicitly designed to address such an ordinal classification scenario. We examined three ordinal strategies [16]:

a) A *'naïve'* transformation of the classification problem into a regression task.

b) A *'naïve'* multi-class decomposition *'One-Vs-One'* (*OvO*) into $\binom{Q}{2}$=3 binary classification problems.

c) An ordinal *'Ordered Partitions'* (*OrdP*) decomposition [44] into Q-1=2 binary problems, formed by contiguous class groupings: (*i*) [Low & Medium] vs. High; and (*ii*) Low vs. [Medium & High].

2.5 Class imbalance & cost-sensitive learning

AI-ML algorithms learn by minimizing an overall loss function averaged across samples in the training dataset. Thus, if certain classes are under-represented in those training data, the algorithms may learn in a biased manner: e.g. prone to suffer marked decays in performance for the minority group(s).

By default, in the typical nominal classification, losses for the different types of errors tend to be assumed all with equal impact. However, this does not hold appropriate with ordinal tasks.

Furthermore, our motivating scenario is inherently cost-sensitive. In clinical practice, the implications for patient safety due to different types of misprognoses should not always be regarded as the same: over-predicting severity may be detrimental to some extent (e.g. for the healthcare providers in their management of personnel and resources), but under-prognoses can imply very serious adverse consequences for those patients receiving suboptimal surveillance or treatment.

To address simultaneously both tasks of learning in a strongly imbalanced class representation, and under a cost-sensitive ordinal scenario, we opted for resampling techniques [45, 46] at the level of data. To tackle imbalance, we first equalized class frequencies in the training dataset using SMOTE oversampling; in particular, its 'Borderline (v1)' variation [47], which focuses on generating synthetic minority samples from instances in the neighborhood of majority ones – hence in danger of being misclassified. (Note that, in those scenarios without effective imputation, random oversampling had to be used instead, since SMOTE does not support missing values).

Afterwards and in order to promote cost-sensitive learning, we made the data imbalanced again, reflecting the ratios of misclassification costs [48, 49]. To do so, we used undersampling with the 'NearMiss (v1)' heuristic [50], which retains instances near the opposite class. Again, in those scenarios without imputation, we used random undersampling).

2.6 AI-ML estimators

The last stage of our proposed pipelines (Figure 1) comprised the AI-ML classification algorithm itself; or in the case of the *'naïve'* ordinal formulation as regression (Section 2.4), the AI-ML regressor. We explored five families of algorithms, to cover a diverse range of AI-ML working principles:

a) Linear methods: Logistic Regression for the two ordinal decompositions based on classification (i.e. *OvO*, *OrdP*), and Ridge for the regression task.

b) Support Vector Machines (SVM) with non-linear kernels; specifically radial basis functions (*'rbf'*).

c) Multi-Layer Perceptron (MLP) architectures, as representatives of shallow artificial neural networks.

d) Random Forests (RF) as *'bagging'* ensembles of decision trees.

e) Histogram-based Gradient Boosting (HGB), within the *'boosting'* ensemble paradigm – Also with the noticeable ability to handle missing values: at each node, HGB learns into which branch they should be routed.

Throughout the training, we automatically tuned each algorithm's key hyperparameters via a cross-validated (CV) grid search (Section 2.11 for implementation details). In particular, we tuned:

a) For the linear models,

a.1) Logistic regression: C_{LR} (inverse of the L^2 regularization strength).

b.2) Ridge: α_{Ridge} (L^2 regularization strength).

b) SVM with *'rbf'* kernel: γ (kernel scale), C_{SVM} (inverse of the L^2 regularization strength).

c) MLP: the number of neurons in the hidden layers, α_{MLP} (L^2 regularization strength).

d) RF: the number of trees in a forest, maximal tree depth, splitting criterion (either Gini impurity or entropy, in classification; mean absolute or squared error, in regression), and the number of candidate features for each split.

e) HGB: η (learning rate), the maximum number of iterations, and maximal tree depth.

Hyperparameter tuning was based on our custom performance score (Section 2.7), defined to account for the trade-offs in class imbalance and cost-sensitive ordinal scenarios.

2.7 Performance metric for model selection

Let $\mathbf{N}$ be the $Q \times Q$ confusion matrix for a certain prediction (here Q=3 target classes). Its $n_{i,j}$ element represents the count of instances assigned to the i-th class but truly belonging to the j-th. Therefore, the total number of ground truth samples in the j-th class is $n_{\bullet j} = \sum_{i=1}^{Q} n_{i,j}$, and the corresponding sensitivity equals ${}^{n_{j,j}}/_{n_{\bullet j}}$. In multi-class imbalanced problems like ours, a widespread metric of performance is the G-mean score (GMS) [49], which equals the geometric mean of per-class sensitivities:

$$GMS \coloneqq \left[\prod_{i=1}^{Q} \frac{n_{j,j}}{n_{\bullet j}} \right]^{1/Q} \tag{1}$$

Besides, let $\mathbf{C}$ be the $Q \times Q$ cost matrix. Its $c_{i,j}$ element quantifies the penalization for predicting a sample as in the i-th class when it truly belongs to the j-th. By convention, $c_{i,i} = 0 \ \forall i \in \{1, \ldots, Q\}$ and $c_{i,j} > 0 \ \forall i \neq j$. Then, known $\mathbf{C}$, the total cost of a predictor with confusion matrix $\mathbf{N}$ equals:

$$\text{Cost}_{\text{Total}} \coloneqq \sum_{i=1}^{Q} \sum_{j=1}^{Q} c_{i,j} n_{i,j} \tag{2}$$

To contextualize this total cost in Eq. (2), we suggest comparing it against a 'dummy' but 'safe' predictor which always outputs the class with minimal total cost:

$$\mathrm{Cost}_{\mathrm{Safe}} := \min_{i\in\{1,\ldots,Q\}} \sum_{j=1}^{Q} c_{i,j} n_{\bullet j} \tag{3}$$

Thus, one can define a cost-based score (CBS), bounded in the $[0,1]$ interval, as follows:

$$\mathrm{CBS} := \max\left\{0, 1 - \frac{\mathrm{Cost}_{\mathrm{Total}}}{\mathrm{Cost}_{\mathrm{Safe}}}\right\} \tag{4}$$

Hence, our cost-sensitive ordinal learning task can be understood as two-sided: *a*) achieving high recognition rates for each and all of the (imbalanced) target classes – i.e. maximizing GMS, and simultaneously *b*) minimizing the overall misclassification cost – i.e. maximizing CBS, thus practical applicability. Inspired by the F_1-score, which in binary classification problems is the harmonic mean between precision and recall, we defined the following custom metric:

$$\mathrm{Score} := 2\frac{\mathrm{GMS}\cdot\mathrm{CBS}}{\mathrm{GMS}+\mathrm{CBS}} \tag{5}$$

i.e. the harmonic mean of GMS and CBS. Here we used this score as our target performance metric –to be maximized– in two contexts: *i*) during AI-ML hyperparameter tuning, and *ii*) for model selection.

2.8 Calibration

When supporting medical decision-making, it is both theoretically sound and practically appropriate [1] to ascertain not only the quality of AI-ML prognoses (Sections 2.7, 2.9) but also the level of uncertainty in prediction. In other words: how close our model estimates class probabilities with respect to the observed relative frequencies. After model selection attending to Eq. (5), our architecture with the best performance was evaluated without and with a final stage of probability calibration [51,52]. Specifically, for calibrating the model we binarized the multi-class problem in a *'One-vs-Rest'* (*OvR*) fashion [53] and employed Platt's sigmoid method [54].

2.9 Final performance evaluation

To benchmark our best model's performance in terms of the success rates attained for our particular cohort and prognostic target, we also calculated the following scoring methods:

a) Pneumonia severity index (PSI) [55], a well-established general-purpose clinical score, which has also been applied in the context of COVID-19 [56].

b) Among the 107 proposals reviewed by Wynants *et al.* [1], the four scores recommended by the authors for hospitalized patients: Xie *et al.* [57], PRIEST [58], ISARIC 4C [59], and Carr *et al.* extension [60] of the NEWS2 score [61]. Note that these four models fulfilled the highest requirements in terms of discrimination ability, calibration, and external validation.

For them, we computed the optimal decision thresholds towards our three-class severity task as proposed in [62]. In the case of [60], we trained their logistic regression from scratch adapting its input: with a binary indicator of whether the patient received supplementary oxygen, instead of the exact flow (not recorded in our study); and with creatinine as a biomarker of kidney function [63], instead of the estimated glomerular filtration rate (eGFR, also not recorded).

Furthermore, we proposed a baseline AI-ML model with a 'standard' design (i.e. non-cost sensitive, non-ordinal): an HGB classifier, whose internal hyperparameters were tuned exactly as in Section 2.6.

As performance metrics, we opted for: accuracy, balanced accuracy (pertinent for class imbalance [49]), area under the receiver operating characteristic (AUC), average precision in the precision-recall curve, Brier score loss (to study accuracy in the probabilistic predictions/calibration), mean absolute error (MAE, a typical metric in ordinal scenarios [16]), GMS (also pertinent for class imbalance [49]), average cost (for cost-sensitive scenarios), and our custom score in Eq. (5).

2.10 Model explainability

Currently, explainable AI-ML is a major topic for both researchers and practitioners, having attracted growing interest. Except for simple algorithms –binary logistic regression or small-sized decision trees, which can be termed as 'transparent' due to their straightforward explainability–, the vast majority of

AI-ML models tend to become very complex internally, with intricate interactions among input variables. Thus, they behave as 'opaque' systems hindering human understanding of AI-ML [64, 65].

To explore and quantify the contribution of each input variable toward severity prognosis, we used Shapley additive explanation techniques (SHAP) [66]. SHAP comprises *post hoc* explainability analyses [65] that compute the marginal impact of a feature (or of a subgroup/'coalition' of them) with respect to the model's outcome. Here we opted for the 'Kernel SHAP' approach, as it is model-agnostic.

2.11 Implementation details

For the sake of reproducibility, we made the source code for our experiments publicly available at our GitHub repository: https://github.com/fegarcia-bcam/CostOrdinalPredict-COVID-19-IEEE-JBHI. Our experiments were run on a supercomputing cluster at the Donostia International Physics Center (DIPC).

We implemented our algorithms based on publicly available Python libraries for AI-ML. We used primarily `scikit-learn` [67]: for the first stages in our pipeline (encoding, scaling, imputation, and most of the feature selection techniques – Figure 1), as well as for the supervised estimators and their calibration. We used `scikit-rebate` [43] for ReliefF feature selection filters, `imbalanced-learn` [68] for data resampling and GMS in Eq. (1), and `shap` [66] for model explainability. We developed in-house implementations to transform the ordinal classification task into regression, *OvO* classification, and *OrdP* problems.

We opted for internal model validation using CV. To prevent information leakage [69], not only the final prediction stage was subjected to CV; but instead the whole pipeline: scaling, imputation, feature selection, ordinal decomposition, and resampling as well (Figure 1).

For model selection, we used 10 independent repetitions of a 5×5-fold nested CV with internal hyperparameter tuning, in order to simultaneously achieve optimal hyperparameters and an unbiased estimation of the models' performance. Samples were doubly stratified into the CV folds both by hospital and by severity level.

We fixed certain pre-processing hyperparameters (Section 2.3): for kNN imputation we opted for k=9 neighbors, and for iterative (MICE) n_f=4 auxiliary features; whereas for ReliefF filters we used k=100 neighbors [42]. In our GitHub repository, the interested reader may find further details about the hyperparameter search spaces employed for the tuning of the five types of AI-ML estimators (Section 2.6).

For the final performance assessment (without and with calibration [51]), we used 20 repetitions of 10-fold CV – again doubly stratified by hospital and by severity. Hyperparameters for the optimal model were set as the most repeated choice during tuning in the model selection stage.

SHAP explainability computations were obtained with a single round of 10-fold CV.

3 Results

3.1 Dataset

With our inclusion/exclusion criteria, a total of n=1548 patients were enrolled: 596 women (38.5%, p<0.001). They were distributed as shown in Table 1, where hospitals have been anonymized. For an exhaustive characterization of our cohort, we kindly refer the reader to our supporting report publicly available at Zenodo: https://doi.org/10.5281/zenodo.7703106. It contains descriptive statistics for all variables –both in overall and by severity–, plots of their univariate distributions, along with hypothesis testing results for differences across groups (χ^2 tests for categorical variables, non-parametric Kruskal-Wallis for continuous) and their effect sizes. This supporting document also depicts maps with the geographical distribution of patients and severities per postcode, socioeconomic information, and the distribution of chronic/acute exposures to pollutants. However, patient confidentiality issues prevented us from making the dataset public.

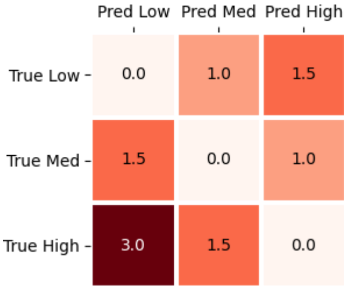
Out of the 106 demographic and clinical variables collected at hospitalization, 14 variables (e.g. ferritin, albumin, bilirubin, or platelets) [42] had to be discarded *a priori* (Figure 1) for not meeting our quality assurance criterion of <60% missing. After 'one-hot' encoding categorical variables, the remaining 92 became 109 features – for further details, please read our Zenodo report. In addition, 7 socioeconomic variables were extracted at each postcode, alongside 32 about chronic and acute exposures to pollutants. Therefore, they totaled d=131 variables before encoding, and d'=148 features afterwards.

Our pulmonologists at the Galdakao-Usansolo University Hospital defined the cost matrix in Figure 2, based on their expert clinical knowledge. As expected for ordinal classification, costs increase monotonically with the difference in the number of levels between the true and the predicted class [16]. The asymmetric structure of this cost matrix encodes the disparate practical consequences of under- and over-prognoses. In all, it entails a trade-off between promoting patient safety and quality of care and constraining the

Table 1: Patient counts and frequencies

		Overall	By severity		
			Low	Medium	High
		n=1548	n=712 (46.0%)	n=238 (15.4%)	n=598 (38.6%)
Hospital	A	358 (23.1%)	205 (57.3%)	36 (10.1%)	117 (32.7%)
	B	380 (24.5%)	229 (60.3%)	50 (13.2%)	101 (26.6%)
	C	438 (28.3%)	119 (27.2%)	59 (13.5%)	260 (59.4%)
	D	372 (24.0%)	159 (42.7%)	93 (25.0%)	120 (32.3%)

number of over-prognoses (to cope with potentially heavy burdens for the clinical staff, in a situation of pandemics).



	Pred Low	Pred Med	Pred High
True Low	0.0	1.0	1.5
True Med	1.5	0.0	1.0
True High	3.0	1.5	0.0

Figure 2: Cost matrix.

3.2 AI-ML model selection

As explained in Sections 2.3 to 2.6 and Figure 1, we explored a comprehensive set of choices with respect to the various stages of our AI-ML pipeline: (*i*) 3 alternatives for imputation; (*ii*) 5 feature selection techniques; (*iii*) 3 ordinal decompositions – regression, *OvO*, *OrdP*; (*iv*) 2 forms of sample rebalancing – equalized to cope with class imbalance, or reimbalanced for cost-sensitive learning; and finally (*v*) 5 families of AI-ML estimators. Not all combinations were feasible, due to the inability of most feature selection algorithms and/or estimators to handle missing data. Thus, we trained and evaluated a total of 260 unique pipeline architectures.

For model selection, Table 2 summarizes their median performance in terms of our custom score in Eq. (5). As highlighted in bold, our best model –i.e. the one with the highest overall score– was an AI-ML pipeline comprising: (*i*) no imputation, (*ii*) no feature selection (i.e. all features fed), (*iii*) *'Ordered Partitions'* for ordinal decomposition, (*iv*) cost-based reimbalance, and (*v*) an HGB classifier handling missing values.

3.3 Final performance evaluation

Having determined our best model (Section 3.2) and its optimal hyperparameters –i.e. the most repeated choice during tuning–, we assessed its overall classification performance via 20 runs of 10-fold CV (Section 2.11), without and with calibration. Figure 3 illustrates its receiver operating characteristic (ROC), precision-recall, and calibration curves. Since such curves are defined for binary problems, we used *'One-vs-Rest'* (*OvR*) transformations as recommended by [70].

In addition, Table 3 contains an exhaustive comparison of the performance results by our best pipeline: against the five models recommended by [1], and against the nominal HGB baseline (Section 2.9). When applied to our dataset with optimal decision thresholds calculated as in [62], these state-of-the-art models were outperformed by both the HGB baseline algorithm and our best cost-sensitive ordinal AI-ML model (also an HGB-powered pipeline).

The HGB baseline behaved equivalently in terms of accuracy, AUC for Low & High severities, precision-recall for High, and MAE (Table 3). Nonetheless, our best model outperformed it in all other performance metrics, particularly in those pertaining class imbalance or cost-sensitiveness: balanced accuracy, GMS, average cost, and our custom score.

Table 2: RESULTS FOR MODEL SELECTION – Median performance score across 10 repetitions of 5×5-fold nested CV with hyperparameter tuning. In bold, our model with the best performance overall.

ESTIMATION	ORDINAL	RESAMPLE	IMPUTATION											
			kNN					*Iterative (MICE)*					*None*	
			FEATURE SELECTION											
			Filt		*Emb*	*Wrap*	*None*	*Filt*		*Emb*	*Wrap*	*None*	*Filt*	*None*
			MI	*RelF*	L^1	*RFE*		*MI*	*RelF*	L^1	*RFE*		*RelF*	
LIN	*Regr*	*Equal*	.119	.118	.146	.104	.268	.083	.088	.131	.082	.253	—	—
	OvO	*Equal*	.320	.301	.336	.339	.370	.296	.289	.338	.340	.345		
		Cost	.356	.343	.369	.367	.345	.329	.348	.380	.366	.367		
	OrdP	*Equal*	.113	.083	.188	.107	.164	0	0	.006	.002	.165		
		Cost	.320	.327	.346	.331	.337	.306	.327	.355	.345	.389		
SVM	*Regr*	*Equal*	.270	.246	.290	.258	.345	.245	.246	.266	.212	.320	—	—
	OvO	*Equal*	.356	.338	.362	.372	.413	.321	.297	.353	.354	.385		
		Cost	.364	.371	.384	.387	.387	.333	.347	.371	.375	.404		
	OrdP	*Equal*	.218	.173	.246	.249	.210	.149	.124	.208	.203	.259		
		Cost	.336	.309	.345	.358	.387	.306	.320	.368	.349	.406		
MLP	*Regr*	*Equal*	.229	.179	.249	.206	.302	.185	.120	.170	.189	.270	—	—
	OvO	*Equal*	.328	.265	.333	.336	.385	.283	.254	.320	.336	.367		
		Cost	.351	.312	.347	.348	.362	.316	.287	.333	.329	.381		
	OrdP	*Equal*	0	0	0	0	.082	0	0	0	0	0		
		Cost	.318	.291	.323	.320	.349	.287	.284	.309	.312	.371		
RF	*Regr*	*Equal*	.208	.200	.208	.162	.212	.161	.155	.170	.134	.167	—	—
	OvO	*Equal*	.371	.371	.386	.390	.399	.350	.340	.372	.374	.394		
		Cost	.384	.365	.398	.392	.371	.358	.364	.384	.384	.385		
	OrdP	*Equal*	0	0	0	0	0	0	0	0	0	0		
		Cost	.388	.369	.396	.393	.414	.368	.364	.391	.391	.424		
HGB	*Regr*	*Equal*	.255	.242	.261	.201	.305	.206	.235	.253	.164	.291	.270	.325
	OvO	*Equal*	.358	.334	.365	.361	.404	.323	.339	.355	.335	.389	.365	.402
		Cost	.358	.352	.375	.370	.359	.333	.340	.376	.360	.388	.405	.437
	OrdP	*Equal*	0	0	0	0	.114	0	0	0	0	.002	.265	.351
		Cost	.344	.348	.369	.369	.396	.337	.349	.363	.354	.412	.380	**.442**

Abbreviations – *kNN*: *k*-nearest neighbors, *Filt*: filter, *MI*: mutual information, *RelF*: ReliefF, *Emb*: embedded, *Wrap*: wrapper, *RFE*: Recursive feature elimination, *LIN*: linear (Ridge for *Regr*, logistic for *OvO* and *OrdP*), *SVM*: support vector machines, *MLP*: multi-layer perceptron, *RF*: random forest, *HGB*: histogram-based gradient boosting, *Regr*: regression, *OvO*: One-vs-One, *OrdP*: ordered partitions.

3.4 Model explainability

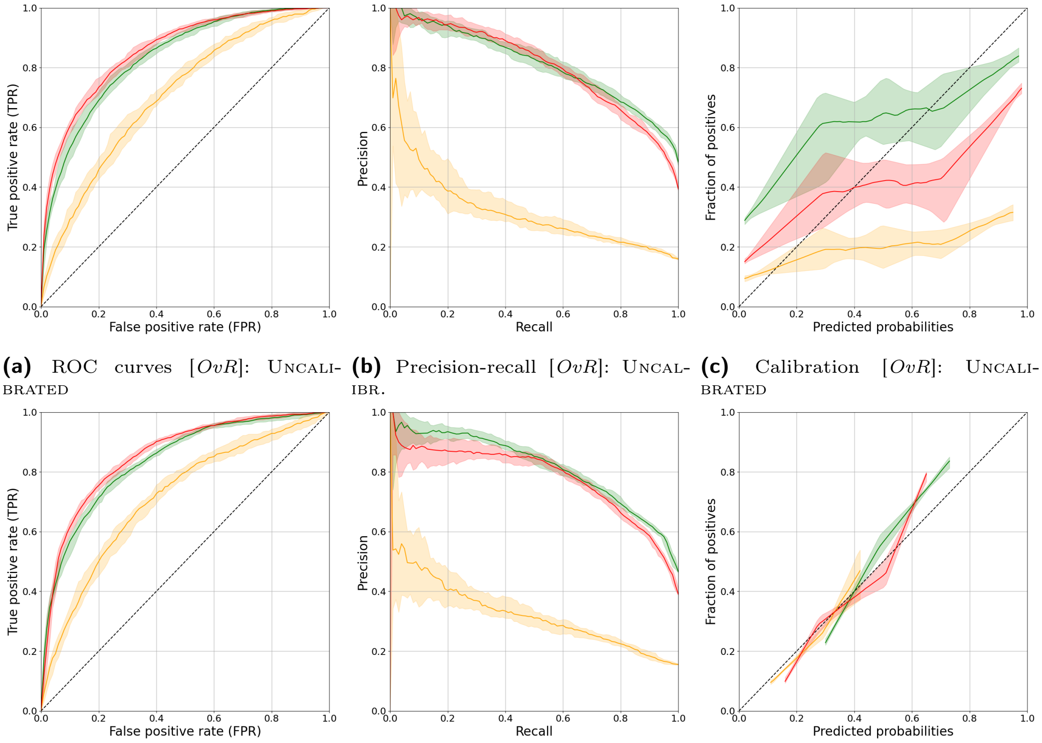
For our best AI-ML pipeline (Section 3.2), Figure 4 depicts the mean absolute value of SHAP explainability magnitudes, per each of the original d=131 variables, and per target severity level – averaged across the n=1548 patients. Notably, our best model working without feature selection (i.e. full input) is consistent with the results displayed in Figure 4, where almost all variables had non-negligible SHAP weights. The only exceptions are #040 (overall COVID-19 symptoms) and #089 (preliminary treatment with Remdesivir during the emergency), but this behavior is easy to understand: only one patient in our cohort was asymptomatic, and only 11 received Remdesivir [Zenodo report].

A dozen variables stood out with the largest SHAP weights: #029 (PSI pneumonia score), #054 (SpO_2/FiO_2) and #054 ($SpO_2/RespiRate$) –both oxygenation biomarkers–, #068 (C-reactive protein, CRP) and #063 (creatinine), #041 (days with symptoms before hospitalization), #092 (emergency treatment with low-molecular-weight heparin), #052 (SpO_2 oxygen saturation), and #067 (lactate dehydrogenase, LDH). Conversely, patients' sex (#001) and age (#002) yielded modest SHAP weights, meaning that our algorithm relied on this information merely to a limited extent.

Socioeconomic factors (#093 to #099) and exposures to air pollution (#100 to #131) also showed non-negligible SHAP weights. Among the latter: #117 ($PM_{2.5}$ acute), #126 (O_3 acute), and #116 (PM_{10} acute).

4 Discussion

This work was based on a multi-centric clinical study for COVID-19, with a broad set of measurements collected to predict the severity of hospitalized SARS-CoV-2 pneumonia patients. Our pneumologists established the relevance of distinguishing three levels of increasing severity, with imbalanced class representation and uneven misclassification penalties. Methodologically, this motivated us to explore AI-ML



(a) ROC curves [*OvR*]: Uncalibrated **(b)** Precision-recall [*OvR*]: Uncalibr. **(c)** Calibration [*OvR*]: Uncalibrated

(d) ROC curves [*OvR*]: Calibrated **(e)** Precision-recall [*OvR*]: Calibr. **(f)** Calibration [*OvR*]: Calibrated

Figure 3: Classification performance by our best model – Without **(top)** and with calibration **(bottom)**. Median and 95% CIs across 20 repetitions of 10-fold CV. Calibration curves **(c, f)** were generated with 5 bins each.
Color coding for ground truth – **Green**: Low, **Orange**: Medium, **Red**: High.

strategies tailored for cost-sensitive ordinal classification. They entail a double novelty with respect to the prevailing algorithmic choices: not only across the comparable COVID-19 literature [1] but in general, with ordinal scenarios.

To exploit this natural order of the ground truth, we explored three decompositions of the ordinal classification problem: (*i*) as a regression task, (*ii*) as a nominal (i.e. non-ordered) *OvO* multi-class classification, and (*iii*) with an explicitly ordinal approach known as *'Ordered Partitions'* (*OrdP*) [44]. In addition, we proposed a custom performance score to guide hyperparameter tuning and model selection – Eq. (5), which encompasses a trade-off between sensitivity to the minority classes, and accounting for the different implications of the various types of over-/under-prognoses. For meaningful comparisons, it used a maximally precautionary (minimal cost) reference – Eq. (3).

We had access to an extensive characterization of patients' status at hospital admission. Nonetheless, there was a high occurrence of missing values, due to the extraordinary burden on the clinical staff during the first wave of COVID-19. This and other peculiarities of our dataset (e.g. variables of different types and scales, feature dimensionality, class imbalance) posed practical challenges which demanded multi-stage pipelines (Figure 1) to address them adequately.

A remarkable strength of our study is having done an exhaustive exploration of AI-ML models tailored for cost-sensitive ordinal tasks: up to 260 different architectures with ordinal decompositions and cost-based resampling. Our pulmonologists incorporated domain knowledge to define the concrete penalizations in our asymmetric cost matrix (Figure 2), but our algorithm is fully generic in this regard: it can be applied straightforwardly with different losses (e.g. penalizing more heavily under-prognoses).

Attending to model selection (Table 2), the order of ground truth classes turned out to be informative, as it was the *OrdP* ordinal decomposition that fostered the top performance. When benchmarked against five well-established scoring methods for pneumonia and COVID-19 hospitalizations (Section 2.9), our our best AI-ML pipeline noticeably outperformed them all across classification metrics (Table 3). In addition, we opted for an extra AI-ML baseline: a 'standard' HGB classifier. This algorithm was comparable in design to our best pipeline, although it followed a nominal approach: neither exploiting ordinal nor cost-

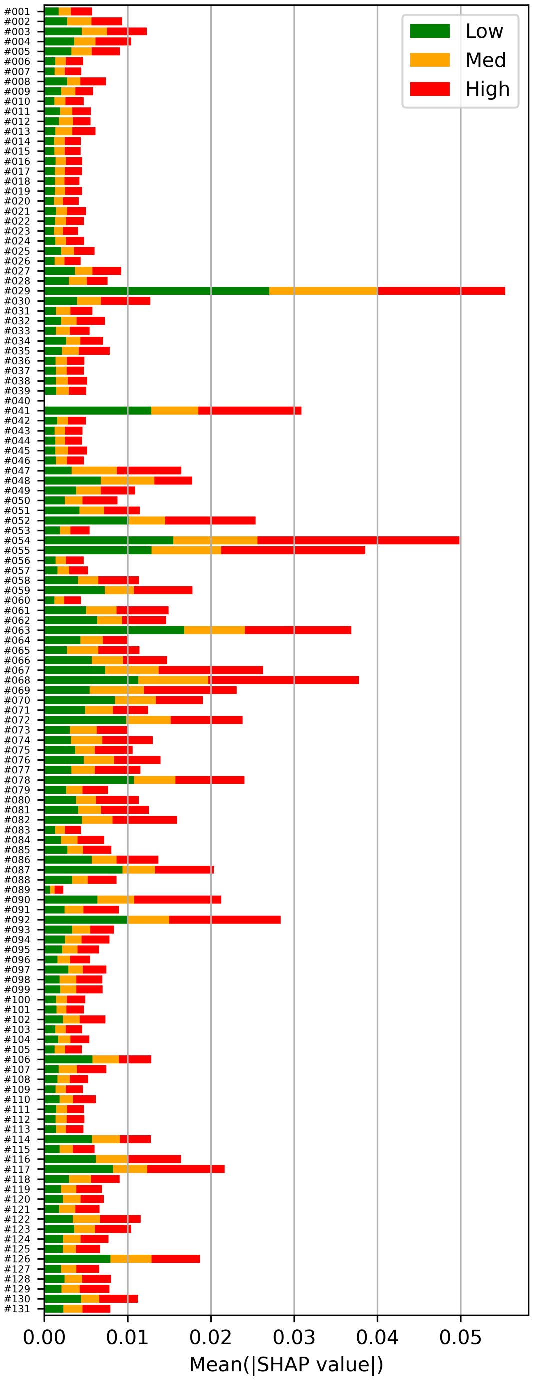


Figure 4: Contributions by variable – SHAP explanation magnitude (average across patients), for each variable in the optimal calibrated model. A detailed list of the variable names and numbers can be found in the appendix materials.

sensitive information. Whereas this HGB baseline yielded equivalent results for certain metrics (accuracy, AUC for Low & High – Table 3), our best model outperformed it in others, noticeably those related to class imbalance or cost-sensitiveness: balanced accuracy, AUC and average precision for Medium (the most minority class), GMS, average misclassification cost, and our custom score. In general, beyond this concrete SARS-CoV-2 pneumonia scenario, we consider that these findings emphasize the appropriateness of tailored ordinal and/or cost-sensitive AI-ML strategies whenever these topic are inherent to the application (as here).

Our best AI-ML pipeline relied on the full set of d' input features. This may indicate that predicting the progression of SARS-CoV-2 pneumonia is a complex task, where many factors play non-negligible roles. The model's behavior in the explainability analyses also reflected this phenomenon, as SHAP showed non-negligible weights for most input variables (Figure 4). SHAP weights were the largest for clinical factors: respiratory status at admission (e.g. PSI score), measurements of oxygenation (SpO_2/FiO_2, $SpO_2/RespiRate$), and biomarkers of inflammation and disease (e.g. CRP, LDH). Yet various features related to pollution exposure also exhibited non-negligible SHAP weights, in line with another study by our COVID-19 & Air Pollution Working Group for this same cohort, in which we found statistically significant effects of various chronic pollutants on the likelihood of death [28].

From a clinical perspective, a limitation of our motivating dataset consists in that it belongs to the first wave of the COVID-19 pandemic in Spain: from February to May 2020. With such a cohort, we aimed at learning patterns from patients who underwent the disease in a situation as homogeneous as possible: regarding the medical knowledge available about COVID-19 and its treatment, and in terms of the burden to the healthcare system. This first-wave situation was detrimental to data collection, which explains –to a major extent– the high occurrence of missing values. We deem it interesting for further research to investigate the algorithmic adaptations to accommodate time-induced distributional shifts [71, 72].

A certain inclusion bias may have been introduced by the admission policies at Hospital C: forced by the unprecedented situation of the pandemic and considering that such an institution had many more ICU beds than other local hospitals, patients who during emergency screenings were assessed as the most fragile or deteriorated, were preferentially referred there. This could explain, to an important extent, the higher rate of severe cases at Hospital C (Table 1).

Finally, calibration –desirable to obtain trustworthy probability estimates (Figure 3c vs. 3f)– appears to have implied some degree of trade-off for various performance scores (Table 3): it improved Brier loss –noticeably, as expected– alongside overall accuracy; but at the expense of balanced accuracy, MAE, GMS, average cost, and our custom score. Thus, practitioners should consider carefully these matters when deciding whether to deploy the model's uncalibrated or calibrated version.

5 Conclusion

We investigated the use of tailored AI-ML strategies for ordinal classification with cost-sensitive learning. The motivating problem was predicting the evolution in severity for hospitalized SARS-CoV-2 pneumonia patients, a scenario in which these 'non-standard' cost-sensitive ordinal topics arise naturally. Despite that, such techniques are often overlooked in the literature, notably in the context of COVID-19.

We conducted exhaustive experiments with 260 different AI-ML architectures, where the top performance was achieved by a model using *'Ordered Partitions'* ordinal decomposition –hence exploiting the information about ground truth class order–, and with cost-based sample reimbalance via resampling for cost-sensitive learning. Furthermore –with accuracy 68.1% [67.3%, 68.8%] (95% CI), balanced accuracy 57.0% [55.6%, 57.9%], and overall AUC 0.802 [0.795, 0.808]–, our best model outperformed five well-established scores for COVID-19 severity, alongside a nominal AI-ML baseline (non-ordinal, cost-insensitive): a *'boosting'* algorithm trained on our dataset. These findings highlight the suitability of exploring beyond the 'standard' nominal techniques when the targeted classification problem is of an ordinal and/or cost-sensitive nature.

Acknowledgments

We would like to acknowledge patients who participated in this research, as well as the staff at the four member hospitals.

COVID-19 & Air Pollution Working Group

La Fe Univ & Polytechnic Hosp: Ana Latorre, Paula González Jiménez, Raul Méndez, Rosario Menéndez. **Cruces Univ Hosp:** Leyre Serrano Fernández, Eva Tabernero Huguet, Luis Alberto Ruiz Iturriaga, Rafael Zalacain Jorge. **Hosp Clínic of Barcelona:** Antoni Torres, Catia Cilloniz. **Galdakao-Usansolo Univ Hosp, Respiratory Service:** Pedro Pablo España Yandiola, Ana Uranga Echeverría, Olaia Bronte Moreno, Isabel Urrutia Landa. **Galdakao-Usansolo Univ Hosp, Research Unit:** Jose María Quintana, Susana García-Gutiérrez, Mónica Nieves Ermecheo, María Gascón Pérez, Ane Villanueva. **BioCruces Bizkaia Health Research Inst:** Mónica Nieves Ermecheo. **Basque Center for Applied Mathematics:** Fernando García-García, Dae-Jin Lee, Joaquín Martínez-Minaya, Miren Hayet-Otero, Inmaculada Arostegui. **IE Univ:** Dae-Jin Lee. **Univ Politècnica de València:** Joaquín Martínez-Minaya. **Tecnalia:** Miren Hayet-Otero. **Univ of the Basque Country:** Miren Hayet-Otero, Inmaculada Arostegui.

Table 3: Comparison of classification performances – For the AI-ML methods, median and 95% CI across 20 repetitions of 10-fold CV. In bold, the best performance for each metric; in italics, the second best.

Metric		**PSI score [55]**	**Xie *et al.* [57]**	**PRIEST [58]**	**ISARIC 4C [59]**	**Adaptation** Carr *et al.* [60]	**Baseline AI-ML [HGB]**		**Our best AI-ML**	
							Uncalibrated	*Calibrated*	*Uncalibrated*	*Calibrated*
Accuracy		.444	.565	.490	.529	.609 [.606, .611]	.680 [.671, .688]	**.681** [.673, .688]	.610 [.595, .629]	**.681** [.673, .688]
Balanced accuracy		.463	.499	.481	.525	.479 [.477, .483]	.568 [.557, .582]	.536 [.530, .541]	**.594** [.580, .615]	*.570* [.556, .579]
AUC	*Overall*	.661	.690	.678	.732	.726 [.724, .727]	.772 [.766, .787]	.769 [.758, .779]	*.799* [.790, .811]	**.802** [.795, .808]
	Low	.710	.731	.728	.775	.756 [.754, .757]	.818 [.808, .830]	**.838** [.832, .844]	.834 [.825, .846]	*.837* [.831, .842]
	Medium	.602	.610	.576	.622	.646 [.641, .650]	.672 [.656, .692]	.614 [.577, .634]	*.704* [.693, .724]	**.714** [.695, .727]
	High	.672	.728	.730	.800	.777 [.774, .777]	.833 [.824, .844]	**.857** [.852, .865]	**.857** [.850, .864]	.855 [.852, .859]
Avg. precis.	*Overall*	.470	.514	.491	.559	.559 [.557, .562]	.614 [.602, .635]	.623 [.614, .634]	**.644** [.633, .663]	*.640* [.627, .650]
	Low	.672	.681	.673	.748	.732 [.729, .734]	.770 [.757, .792]	.803 [.796, .812]	*.811* [.798, .827]	**.815** [.804, .824]
	Medium	.196	.215	.189	.209	.250 [.241, .255]	.305 [.291, .339]	.269 [.241, .292]	*.318* [.290, .354]	**.326** [.303, .344]
	High	.542	.647	.611	.718	.696 [.692, .698]	.765 [.750, .787]	*.800* [.793, .815]	**.801** [.794, .819]	.777 [.767, .792]
Brier score (BS) loss		.559	.535	.537	.493	.510 [.509, .512]	*.484* [.454, .503]	.473 [.469, .476]	.650 [.619, .674]	**.451** [.447, .456]
Mean absolute error (MAE)		.685	.641	.661	.612	.627 [.622, .632]	**.478** [.463, .494]	.485 [.470, .499]	.501 [.480, .526]	*.479* [.467, .492]
Geometric mean score (GMS) [Eq. (1)]		.449	.463	.472	.496	.162 [.129, .186]	*.455* [.431, .492]	.000 [.000, .073]	**.583** [.569, .606]	.449 [.415, .477]
Average cost $1/n \cdot Cost_{Tot}$ [Eq. (2)]		.843	.747	.704	.610	.738 [.732, .745]	.577 [.562, .593]	.566 [.551, .581]	**.527** [.505, .556]	*.546* [.534, .569]
Custom score [Eq. (5)]		.001	.184	.244	.355	.139 [.123, .154]	.374 [.352, .396]	.137 [.135, .170]	**.455** [.426, .483]	*.391* [.366, .414]

References

[1] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

[2] R. Vaishya *et al.*, "Artificial Intelligence (AI) applications for COVID-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 337–339, 2020.

[3] J. S. Suri *et al.*, "Systematic review of artificial intelligence in acute respiratory distress syndrome for COVID-19 lung patients: A biomedical imaging perspective," *IEEE J Biomed Health Inf*, vol. 25, no. 11, pp. 4128–4139, 2021.

[4] J. Musulin *et al.*, "Application of artificial intelligence-based regression methods in the problem of COVID-19 spread prediction: A systematic review," *Int J Environ Res Public Health*, vol. 18, no. 8, 2021.

[5] G. Lombardo *et al.*, "Fine-grained agent-based modeling to predict COVID-19 spreading and effect of policies in large-scale scenarios," *IEEE J Biomed Health Inf*, vol. 26, no. 5, pp. 2052–2062, 2022.

[6] Y. Liu *et al.*, "A COVID-19 risk assessment decision support system for general practitioners: Design and development study," *J Med Internet Res*, vol. 22, no. 6, p. e19786, 2020.

[7] Y. Ye *et al.*, "α-Satellite: An AI-driven system and benchmark datasets for dynamic COVID-19 risk assessment in the United States," *IEEE J Biomed Health Inf*, vol. 24, no. 10, pp. 2755–2764, 2020.

[8] L. Jehi *et al.*, "Individualizing risk prediction for positive coronavirus disease 2019 testing: Results from 11,672 patients," *Chest*, vol. 158, no. 4, pp. 1364–1375, 2020.

[9] W. Shi *et al.*, "COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks," *IEEE J Biomed Health Inf*, vol. 25, no. 7, pp. 2376–2387, 2021.

[10] S. R. Knight *et al.*, "Risk stratification of patients admitted to hospital with COVID-19 using the ISARIC WHO clinical characterisation protocol: Development and validation of the 4C mortality score," *BMJ*, vol. 370, 2020.

[11] M. Luo *et al.*, "IL-6 and CD8+ T cell counts combined are an early predictor of in-hospital mortality of patients with COVID-19," *JCI Insight*, vol. 5, no. 13, 2020.

[12] L. Meng *et al.*, "A deep learning prognosis model help alert for COVID-19 patients at high-risk of death: A multi-center study," *IEEE J Biomed Health Inf*, vol. 24, no. 12, pp. 3576–3584, 2020.

[13] S. Tabik *et al.*, "COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images," *IEEE J Biomed Health Inf*, vol. 24, no. 12, pp. 3595–3605, 2020.

[14] M. van der Schaar *et al.*, "How artificial intelligence and machine learning can help healthcare systems respond to COVID-19," *Machine Learning*, vol. 110, no. 1, pp. 1–14, 2021.

[15] F.-Y. Cheng *et al.*, "Using machine learning to predict ICU transfer in hospitalized COVID-19 patients," *Journal of Clinical Medicine*, vol. 9, no. 6, 2020.

[16] P. A. Gutiérrez *et al.*, "Ordinal regression methods: Survey and experimental study," *IEEE Trans Knowl Data Eng*, vol. 28, no. 1, pp. 127–146, 2016.

[17] A. Fernández *et al.*, *Learning from imbalanced data sets*. Springer, 2018, ch. Cost-sensitive learning, pp. 63–78.

[18] B. G. Pijls *et al.*, "Demographic risk factors for COVID-19 infection, severity, ICU admission and death: A meta-analysis of 59 studies," *BMJ Open*, vol. 11, no. 1, 2021.

[19] R. B. Hawkins *et al.*, "Socio-economic status and COVID-19–related cases and fatalities," *Public Health*, vol. 189, pp. 129–134, 2020.

[20] P. Congdon, "COVID-19 mortality in English neighborhoods: The relative role of socioeconomic and environmental factors," *J*, vol. 4, no. 2, pp. 131–146, 2021.

[21] C. Copat *et al.*, "The role of air pollution (PM and NO_2) in COVID-19 spread and lethality: A systematic review," *Environ Res*, vol. 191, p. 110129, 2020.

[22] N. Ali and F. Islam, "The effects of air pollution on COVID-19 infection and mortality - A review on recent evidence," *Front Public Health*, vol. 8, 2020.

[23] Z. J. Andersen *et al.*, "Air pollution and COVID-19: clearing the air and charting a post-pandemic course: a joint workshop report of ERS, ISEE, HEI and WHO," *Eur Respir J*, vol. 58, no. 2, p. 2101063, 2021.

[24] T. Bourdrel *et al.*, "The impact of outdoor air pollution on COVID-19: a review of evidence from in vitro, animal, and human studies," *European Respiratory Review*, vol. 30, no. 159, p. 200242, 2021.

[25] A. Frontera *et al.*, "Severe air pollution links to higher mortality in COVID-19 patients: The "double-hit" hypothesis," *J Infect*, vol. 81, no. 2, pp. 255–259, 2020.

[26] B. Wachtler *et al.*, "Socioeconomic inequalities and COVID-19, A review of the current international literature," *J Health Monit*, vol. 5, no. Suppl 7, pp. 3–17, 2020.

[27] S. Khalatbari-Soltani *et al.*, "Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards," *J Epidemiol Community Health*, vol. 74, no. 8, pp. 620–623, 2020.

[28] O. Bronte *et al.*, "Impact of outdoor air pollution on severity and mortality in COVID-19 pneumonia," *Sci Total Environ*, vol. 894, p. 164877, 2023.

[29] F. Cabitza and A. Campagner, “The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies,” *Int J Med Informatics*, vol. 153, p. 104510, 2021.

[30] INE Spanish National Statistics Institute, “Household income distribution atlas,” 2019. [Online]. Available: https://www.ine.es/dynt3/inebase/en/index.htm?padre=7132

[31] Basque Network for the Surveillance of Air Quality, “Air quality measurements in the Basque Country,” 2019. [Online]. Available: https://www.opendata.euskadi.eus/catalogo/-/calidad-aire-en-euskadi-2019

[32] ——, “Air quality measurements in the Basque Country,” 2020. [Online]. Available: https://www.opendata.euskadi.eus/catalogo/-/calidad-aire-en-euskadi-2020

[33] Catalan Network for the Monitoring and Prediction of Air Pollution, “Air quality measurements in Catalonia.” [Online]. Available: https://analisi.transparenciacatalunya.cat/es/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom%C3%A0t/tasf-thgu

[34] Valencian Network for the Monitoring and Surveillance of Air Pollution, “Air quality measurements in the Valencian Community.” [Online]. Available: https://agroambient.gva.es/es/web/calidad-ambiental/datos-historicos

[35] N. Umlauf *et al.*, “BAMLSS: Bayesian additive models for location, scale, and shape (and beyond),” *Journal of Computational and Graphical Statistics*, vol. 27, no. 3, pp. 612–627, 2018.

[36] H. D. Alas *et al.*, “Pedestrian exposure to black carbon and $PM_{2.5}$ emissions in urban hot spots: new findings using mobile measurement techniques and flexible Bayesian regression models,” *J Exposure Sci Environ Epidemiol*, 2021.

[37] O. Troyanskaya *et al.*, “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[38] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: A critical evaluation,” *BMC Med Inform Decis Mak*, vol. 16, no. S3, 2016.

[39] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in R,” *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[40] J. N. Wulff and L. Ejlskov, “Multiple imputation by chained equations in praxis: Guidelines and review,” *Electron J Bus Res Methods*, vol. 15, no. 1, pp. 41–56, 2017.

[41] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[42] M. Hayet-Otero *et al.*, “Extracting relevant predictive variables for COVID-19 severity prognosis: An exhaustive comparison of feature selection techniques,” *PLoS One*, vol. 18, no. 4, p. e0284150, 2023.

[43] R. J. Urbanowicz *et al.*, “Benchmarking Relief-based feature selection methods for bioinformatics data mining,” *J Biomed Inform*, vol. 85, pp. 168–188, 2018.

[44] E. Frank and M. Hall, “A simple approach to ordinal classification,” in *Machine Learning: ECML*, 2001, pp. 145–156.

[45] V. López *et al.*, “Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics,” *Expert Syst Appl*, vol. 39, no. 7, pp. 6585–6608, 2012.

[46] H. He and Y. Ma, *Imbalanced learning: Foundations, algorithms, and applications.* Wiley, 2013.

[47] H. Han *et al.*, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” *Lect Notes Comput Sci*, vol. 3644, pp. 878–887, 2005.

[48] C. Elkan, “The foundations of cost-sensitive learning,” in *17th International Joint Conference on Artificial Intelligence*, 2001.

[49] R. Barandela *et al.*, “Strategies for learning in class imbalance problems,” *Pattern Recognit*, vol. 36, no. 3, pp. 849–851, 2003.

[50] I. Mani, “kNN approach to unbalanced data distributions: A case study involving information extraction,” in *Workshop on Learning from Imbalanced Datasets*, 2003.

[51] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *22nd International Conference on Machine Learning*, 2005, pp. 625–632.

[52] B. Zadrozny and C. Elkan, “Obtaining calibrated probability estimates from decision trees and naïve Bayesian classifiers,” in *18th International Conference on Machine Learning*, vol. 1. Citeseer, 2001, pp. 609–616.

[53] ——, “Transforming classifier scores into accurate multiclass probability estimates,” in *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 694–699.

[54] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[55] M. J. Fine *et al.*, “A prediction rule to identify low-risk patients with community-acquired pneumonia,” *N Engl J Med*, vol. 336, no. 4, pp. 243–250, 1997.

[56] A. Artero *et al.*, “Severity scores in covid-19 pneumonia: a multicenter, retrospective, cohort study,” *J Gen Intern Med*, vol. 36, no. 5, pp. 1338–1345, 2021.

[57] J. Xie *et al.*, "Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19," medRxiv, Tech. Rep., 2020.

[58] S. Goodacre *et al.*, "Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: The PRIEST observational cohort study," *PLoS One*, vol. 16, no. 1, p. e0245840, 2021.

[59] R. K. Gupta *et al.*, "Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study," *Lancet Respir Med*, vol. 9, no. 4, pp. 349–359, 2021.

[60] E. Carr *et al.*, "Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study," *BMC Med*, vol. 19, no. 1, 2021.

[61] G. B. Smith *et al.*, "The National Early Warning Score 2 (NEWS2)," *Clin Med*, vol. 19, no. 3, pp. 260–260, 2019.

[62] C. T. Nakas *et al.*, "Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index," *Stat Med*, vol. 29, no. 28, pp. 2946–2955, 2010.

[63] W. R. Zhang and C. R. Parikh, "Biomarkers of acute and chronic kidney disease," *Annu Rev Physiol*, vol. 81, no. 1, pp. 309–333, 2019.

[64] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf Fusion*, vol. 76, pp. 89–106, 2021.

[65] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Front Big Data*, vol. 4, 2021.

[66] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Adv Neural Inf Process Syst*, 2017, pp. 4766–4775.

[67] F. Pedregosa *et al.*, "scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[68] G. Lemaître *et al.*, "imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[69] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[70] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, 2006.

[71] Y. Dendramis *et al.*, "Estimation of time-varying covariance matrices for large datasets," *Econometric Theory*, vol. 37, no. 6, pp. 1100–1134, 2021.

[72] C. Huyen, *Designing machine learning systems.* O'Reilly Media, 2022.