

Original Research

Received 05 June 2026

Revised 19 July 2026

Accepted 02 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (15s): 269–284

KEYWORDS:

COVID-19; intensive-care prediction; clinical diagnosis; ensemble learning; data leakage; explainable artificial intelligence; SHAP; confidence intervals; Cohen's kappa

Leakage-Free, Explainable Ensemble Machine Learning for COVID-19 Clinical Risk Stratification: A Two-Task Study of Intensive-Care Prediction and Blood-Marker Diagnosis

Shaba Parveen Khan¹, Sanjay Jain²

¹Research Scholar, Amity University Jaipur Rajasthan, India, Department of Computer Science and Engineering, ershaba@gmail.com

²Assistant Professor Amity University Jaipur Rajasthan, India, Department of Computer science and engineering , sjain@jpr.amity.edu , Corresponding Author: Shaba Parveen Khan*

ABSTRACT: Many of the clinical machine-learning struggles on COVID-19 have two common flaws. Models are often described to perform well, because data is leaked into them, and they are "black boxes" designed that a clinician cannot be requested to examine. We do both at once by executing a single, clean pipeline on two distinct clinical tasks on distinct, public datasets both at once.

The first task foretells admission to intensive care. It is based on 385 patients from a Brazilian tertiary hospital, and we only accept measurements the first two hours after arrival, exclude the patients who are already in the ICU during this period and set a very strict definition of what we consider to be a transfer, which is from the outside to the inside of the hospital. The second was a diagnostic task, which involved identifying SARS-CoV-2 infection from a common blood test called a complete blood count (CBC) in a separate sample of 5,644 individuals.

Six ensembles were tested under the same conditions on the same data (Random Forest, Extremely Randomised Trees, Gradient Boosting, AdaBoost, LightGBM, XGBoost), using stratified splitting, 5-fold cross validation, and metrics that take into account class imbalance. We applied Permutation Importance and SHAP to open up the models that we obtained. Additionally, held-out estimates per class are reported along with 95% Wilson score confidence intervals, and inter-rater agreement style (Cohen's kappa) is compared to the Landis-Koch scale, which includes an explicit measure of statistical uncertainty that is not expressible from raw point estimates alone.

Boosting-family models, AdaBoost and Gradient Boosting, performed better than both bagging-family models on both tasks, out-performing the Random Forest and Extra-Trees methods. In the ICU problem the best models obtained very high sensitivities, based primarily on age percentile, neutrophil and lymphocyte counts, and renal/cardiovascular markers. AUC values ranged from 0.74 to 0.80 for diagnosis, and the eosinophil, leukocyte, and platelet counts contributed most to the AUC. It should be emphasized that features revealed by permutation importance are similar to those reported in the clinical literature.

The bigger depiction: When the leakage is known, when a post-hoc explainability report is provided, when a formal statistical-uncertainty report is available, and when the implementation is complexity-aware, a prediction has a firmer and more clinically usable, and more reviewable, footing than can be established by a headline accuracy figure. But,

with small sample sizes and the unevenness of the data that remains in each class, it is important to be cautious before generalizing these results.

1. INTRODUCTION

In the case of a respiratory infection epidemic, the response to this seemingly simple question is the response to life and death here: who will deteriorate? Having the answer early allows for the beds, ventilators and staff to be prepped before the peak and a slipping patient can be escalated while it is still helpful. Capacity for getting it late and critical care is used reactively – typically after it is too late. This shifted from a seasonal concern to a global crisis due to COVID-19, and the number of machine-learning studies predicting severe outcomes and the need for intensive care units to routinely collected clinical data has surged.

That literature has two issues that need to be addressed. The first is technical and is more common than it should be; data leakage. Train a model to predict a patient's admission to the ICU, and then show it some data collected during or after the patient's admission, and the model learns to identify an ICU patient that is already critically ill not that they will end up being admitted to the ICU. The accuracy is on paper, very high. In practice it doesn't work since at the moment of admission the leaked signal does not yet exist. Without a proper containment area for measurement and results, there will be an inevitable form of hard cutoff between function and function outcome. The second is the issue of Trust. The best models on top of the leaderboard on tabular clinical data are typically gradient-boosting ensembles, but they're also the most opaque and a doctor not knowing why a prediction was made will not, sensibly, follow it.

This uncertainty is the reason for the growing shift in the clinical prediction domain toward explainable AI (XAI). Now they're considered a staple and used in various methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which enable a clinician to understand what features influence a prediction, both on a group level and for an individual patient. With an explanation, a model has a basis of physiologically sensible variables: lymphopenia, increased inflammatory markers, increased age. Then, a clinical defense of the model becomes much easier. There's a regulatory aspect, as well. Finally, the emerging standards for clinical decision support, transparency, auditability and validation in the intended population make interpretability a responsibility of the deployment of predictive tools rather than an extra.

In this context, our contribution is quite limited. The algorithm we don't propose is new. Rather, we develop and evaluate a predictive task for intensive care, which is strictly leakage-free, use this task to compare the main ensemble families on an equal footing, provide explanations for them using two separate methods, and beyond merely comparing the points formalize the statistical uncertainty of each per-class figure and each chance-corrected agreement measure against a pre-established scale and rigorously read the numbers with an awareness of how much confidence they can carry.

Major Contributions of this paper are:

- We set up an early, leakage-free risk-stratification task: During the first 2 hours after admission, the inputs are limited to the first 2 hours, any patient already in ICU during the first 2 hours are excluded, and the output of this task will be a subsequent ICU admission.
- We scored the six ensemble classifiers with metrics that were appropriate for an imbalanced binary outcome, and cross-validated them over five folds, and so we measure the stability of the ensemble, not just one split.
- We used permutation importance and SHAP to read the models, revealing a clinically coherent set of predictors which leave the model unchanged where it agreed and remained candid about what the model didn't and did predict.
- We repeat the whole approach on a second, independent dataset, but for a different clinical prediction problem (the example of COVID-19 diagnosis from a routine blood count), demonstrating that the method is transferrable not just between datasets, but between different clinical prediction problems.
- For each held-out per-class estimate we provide a mathematical description of all models and metrics applied, we quantify the complexity of the pipeline, and we estimate a 95% confidence interval for each estimate with a number based on the exact count of the samples, with an agreement-strength benchmark attached to each number, so that the precision implied by each reported number is explicitly stated and not assumed.

A paper is organized as: The survey of recent work related to this work is provided in Section 2. In Section 3, the data, task construction without leakage, mathematical formulation of the models and metrics, evaluation protocol and implementation environment are presented. The comparative results, explainability analysis and the added uncertainty and agreement-benchmarking analyses are presented in Section 4. The section is concluded with limits and suggestions for further research.

2. RELATED WORK

2.1 Comparative classifier studies

Tree-based ensembles work well on tabular clinical data. Several different frameworks were compared in one general study of the COVID-19 disease-severity and mortality prediction: Random Forest combined with feature selection was found to be accurate, and blending blood-test and vital-sign features improved the mortality prediction [1]. A four-classifier comparison using a combination of hyperparameter tuning, calibration, DCA and SHAP has been performed at population scale to predict hospitalization on top of the best model [4] and a BMC Public Health study has created and compared severity-risk models on routine data [5]. In other areas, however, clinical blood-test data have proven to be useful in predicting the severity and have been validated in an external cohort [8].

2.2 Gradient boosting in clinical risk prediction

Since COVID-19, gradient boosting methods have become widely used for assessing and predicting risk. LightGBM, CatBoost and XGBoost have been compared on clinical outcomes and inflammatory markers have been marked as important [9, 11] repeatedly by SHAP. LightGBM models have been reported for prediction of coronary heart disease [9] and, externally validated, colorectal cancer risk [11] and hybrid boosting for transferable diabetes risk stratification continues to outperform simpler baselines [12]. Comparisons between head-to-head XGBoost and LightGBM have also been done in the cancer-risk prediction field [13] along with feature-importance analysis.

It is important to understand that there is a problem of explainability, and that the leakage problem arises from this. By now XAI is standard equipment. SHAP is used to rank predictive attributes for severity groups (alongside LIME) [7] and to emerge high impact features like age and neutrophils (in a multi-modality model) [3, 2]. Predictive modelling workflows for the care setting and target population are poorly described in many published COVID-19 prognostic models, and many of these models are at high risk or unclear risk of bias [10] precisely the type of failure that leakage would create. Our study takes a shot at that critique, directly exposing the temporal distance between feature and outcome.

This study is compared to representative recent work in Table 1. A six-model comparison, a task that could be guaranteed to be leakage-free, dual explainability, and a stability analysis.

Table 1. Recent studies (2023–2026).

Study	Models	Leakage control	Explainability	Stability
Disease severity ML [1]	7 classifiers	Not stated	No	CV
Hospitalization, large cohort [4]	4 classifiers	Not stated	SHAP	Calibration
Blood-test severity [8]	Several	Partial	Limited	External
SHAP+LIME severity [7]	Several	Not stated	SHAP+LIME	No
Proposed study	6 ensembles	Explicit (first-window)	Permutation+SHAP	5-fold CV + 95% CI

This study finds the gap as three things are evident in the studies presented in Table 1: Point estimates of accuracy, F1 or AUC are reported, but the sampling uncertainty surrounding the point estimates is rarely mentioned and seldom is Cohen's kappa measured relative to a standard interpretive scale. It matters on small and imbalanced test-partitions which are typical of single-

centre COVID-19 cohorts, with a third of probability scale being covered by a 95% interval for a recall of 0.76 built on 17 events. In Section 3 and Section 4, we use the Landis–Koch kappa scale [22] and the Wilson score interval [23].

3. MATERIALS AND METHODS

3.0 Background: ensemble learning and explainability

Ensemble methods are used to combine several decision trees to overcome the instability and low accuracy of a single decision tree. The main difference between the two dominant strategies is how they do the stacking. The trees are grown independently on bootstrap resamples and the votes are averaged in the bagging process, which reduces variance and yields robust models like Random Forest [15] and Extremely Randomised Trees (Extra-Trees) [16]. Boosting is growing trees sequentially, each one correcting the errors of its past; this is usually reducing bias and looking to squeeze as much accuracy out of structured tabular data, at the expense of being more sensitive to noise and hyperparameters. AdaBoost [18] re-weights the misclassified samples. Each new tree is trained on the gradient of the loss, Gradient Boosting [17] fits each new tree to the gradient of the loss. The LightGBM and XGBoost are the tuned versions of gradient boosting that include regularisation, histogram-based split finding, and native class weighting. The natural design is to pitted both families against each other, because boosting is at its best when applied to clinical tabular data.

This type of accuracy is achieved at the price of interpretability. Permutation importance considers the fitted model as a black box and asks how much worse the model performs when the feature values are shuffled; if the performance is significantly worse then it is a good sign that it is really important for the model. Instead, SHAP [21] adopts a game-theoretic approach and distributes an additive score to each feature on each individual prediction, which can then be aggregated into a global score. The two are answering slightly different questions – how much performance depends upon a feature, versus how much a feature stresses each prediction – so the places where they agree deserve weight.

3.1 Datasets

Two independent and public COVID-19 cohorts, both from Brazilian centres, are used. The first one, for an intensive-care task, was retrieved from Hospital Sírio-Libanês / Hospital Israelita Albert Einstein and published on Kaggle [14]. It contains 1,925 records on 385 distinct patients over five consecutive time periods (0-2, 2-4, 4-6, 6-12, and >12 hours) that include demographic data, groups of pre-existing disease, and summarized laboratory and vital signs data, with the outcome being admission to the intensive care unit (ICU). The second, also for diagnosis, is the dataset of Hospital Israelita Albert Einstein, which includes a total of 5,644 patients who went to the ER of the hospital, each with a routine complete blood count and a result of the test for the presence of SARS-CoV-2 RT-PCR. There are many common haematologic features to both cohorts, but they enable different prediction tasks, allowing one methodology to be tested on two different prediction tasks.

3.2 Leakage-free task construction

This is the serious and well documented problem of predicting ICU admission from this data set: the model will learn to detect patients who are already very sick, and thus will never be able to perform well when used in real time admission. We make the task to close that door by explicitly separating them in time. First, it puts the outcome at the patient level and at any window of time, the patient will be admitted to the ICU. Second, information is only available from the first window (0-2 hours after admission) and this is information that is actually available early. Third, the 32 patients already in the intensive care unit (ICU) at that first time window are excluded because they are already in the “end point” and the outcome is not a future event. What is left is a clean cohort of 353 patients: 190 patients admitted to the hospital without ever being hospitalized in an ICU, and 163 patients who were admitted to an ICU later in their hospital stay — a close-to-balanced binary problem. This is the key methodological protection used in this study and it addresses the issue of risk of bias raised in previous prognostic models [10].

3.3 Features and pre-processing

Nineteen clinically interpretable variables were retained in the first window: age percentile and an age-above-65 flag, gender, hypertension and immunocompromised, the vital signs (spO₂, respiratory rate, heart rate, temperature, SBP, DBP) and the routine laboratory markers (CRP, lactate, lymphocytes, neutrophils, creatinine, urea, hemoglobin, platelets). There were missing values in the early-window laboratory panels for some patients, which were filled with the median value predicted by the training-set panel using the same data that was used for the modelling. Any information from the test set was not leaked in to the training set and all the features were standardized. A fixed random seed ran throughout, for reproducibility.

Formally, if feature j is missing at position i , then $\hat{x}_{ij}$ is the median of the values at positions k , $k \in$ training fold, and the values retained at positions k , $k \in$ training fold are then standardized as $z_{ij} = (x_{ij} - \mu_j)/\sigma_j$, where μ_j and σ_j are the mean and standard deviation of the training fold respectively. In addition to the temporal leakage control described in Section 3.2, the pre-processing itself is also free of leakage because these statistics are calculated using the held-out partition for training, and are used as given, without re-calculation on the held-out partition for validation or test.

The cohort was split into training (80%), validation (10%) and test (10%) sets in a stratified fashion with proportions maintained in each set. The training subset was used for fitting and five-fold cross-validation, the validation subset was used for monitoring, and the test subset was only touched when obtaining the final estimate.

3.4 Ensemble models

On the side of ensemble classifiers, 6 classifiers were specified under identical conditions: Random Forest and Extra-Trees on the bagging side, Gradient Boosting, AdaBoost, LightGBM and XGBoost on the boosting side. Where possible class weighting (or, where not native, balanced sample weighting) was introduced to counterbalance the remaining imbalance. Each classifier used the same imputation and standardization process, ensuring a like for like comparison. The key hyperparameters are listed in Table 2, and were set to reasonable and consistent values between all runs to prevent the gaps in performance from being due to suboptimal tuning. The reproduction in Section 4 was performed in an off-line setting, and only the four models directly supported by scikit-learn (Random Forest, Extra-Trees, Gradient Boosting, AdaBoost) were able to be fit, with the other models (LightGBM, XGBoost) having their hyperparameters described below as the specification that a future internet-connected reproduction should follow.

Table 2. Principal hyperparameters of the six ensemble classifiers (four fit and evaluated in this reproduction; LightGBM/XGBoost specified for future reproduction).

Model	Estimators	Key settings
Random Forest	400	balanced class weight; full-depth trees
Extra-Trees	400	balanced class weight; randomised splits
Gradient Boosting	300	learning rate 0.05; max depth 3
AdaBoost	300	learning rate 0.5
LightGBM*	400	learning rate 0.05; 31 leaves; balanced
XGBoost*	400	learning rate 0.05; max depth 4; subsample 0.9

*Not fit in this reproduction (library unavailable offline); hyperparameters given for future reproduction.

A formal definition of a model of computation and an understanding of related subject areas are developed. Formal model definitions and computational complexity understood.

Let n be the number of patients, p be the number of features and B/M be the number of trees/rounds. Random Forest [15] selects B bootstrap samples, builds an unpruned tree h_b for each of them, and classifies on the basis of hard voting,

$$\hat{y}(x) = \text{mode}\{h^1(x), \dots, h_{B(x)}(x)\} \quad \text{Eq. (1).}$$

In addition to this, Extra-Trees [16] introduces randomized split values, sacrificing a small amount of bias for reduced variance. The cost of training is $O(B \cdot n \cdot \log n \cdot \sqrt{p})$ for both.

Gradient Boosting [17] constructs the model stage by stage using the negative gradient of the loss with respect to the previous model F_{m-1} and adds it with learning rate v ,

$$F_m(x) = F_{m-1}(x) + v \cdot h_m(x) \quad \text{Eq. (2).}$$

Boosting's original discrete form is AdaBoost [18] whose samples are re-weighted $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot \mathbb{1}(y_i \neq h_m(x_i))]$, where $\alpha_m = \frac{1}{2} \ln[(1 - e_m)/e_m]$ is determined by the weighted error rate e_m — Eq. (3). Unlike [17] and [18] which regularised the objective implicitly, XGBoost [19] regularises it explicitly, and trains each round with a second order (Newton) loss approximation: $\ell(\varphi) = \sum_i \ell(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$, with $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ penalising the number of leaves T and the leaf weights w — Eq. (4). LightGBM [20] effectively achieves the same objective of the boosted trees, but using histogram-binned, $O(n)$ split search instead of $O(n \cdot \log n)$ sorting, gradient-based one-side sampling, and leaf-wise tree growth, and in fewer splits to obtain a certain reduction in loss. On this study, the sample size n (up to 385) is not the limitation, but the histogram-based boosters, XGBoost and LightGBM, are faster than exact, sort-based boosters (Gradient Boosting and AdaBoost), which are $O(M \cdot n \cdot p)$ per round.

3.5 Evaluation metrics

In the case of a binary outcome, and in the case of the ICU class in particular, we report the precision, the recall and their harmonic mean (F1), which are all calculated from true positives (TP), false positives (FP), and false negatives (FN):

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}), \quad \text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (5)$$

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (6)$$

For both classes we report macro-averaged F1 (unweighted mean F1 scores per class, $\text{F1}_{\text{macro}} = (1/C) \cdot \sum_c \text{F1}_c$) and weighted averaged F1 (weighted mean F1 scores per class according to the class's support n_c , $\text{F1}_{\text{weighted}} = \sum_c (n_c/n) \cdot \text{F1}_c$). We include overall accuracy, Cohen's kappa and the area under the receiver operating characteristic curve (AUC).

Cohen's kappa adjusts the agreement p_o (accuracy) observed in the data, for the agreement p_e (expected by chance) based on the marginal class proportions:

$$\kappa = (p_o - p_e) / (1 - p_e) \quad (7)$$

The AUC is the probability that a randomly selected positive case is scored by the model higher than a randomly selected negative case, which equals to the Mann–Whitney U statistic minus a constant (which value depends on the problem being solved). On the ICU, class of the recall is most important; a missed patient in the ICU is the most expensive error of all. In addition, the mean and the standard deviation of macro-F1 are reported for the training data over five stratified cross-validation folds, in order to describe stability, not just one lucky split.

3.6 Explainability

Two complimentary methods were used for the task. The feature importance was calculated on the held_out test set: For each feature j , the importance score is computed as the mean drop in macro-F1 over the K repetitions, where the column of the test set corresponding to feature j is randomly permuted over the K repetition.

$$\text{PI}_j = (1/K) \cdot \sum_{k=1}^K [\text{Macro-F1}(y, \hat{y}) - \text{Macro-F1}(y, \hat{y}_{\pi_k(j)})] \quad (8)$$

A high positive PI_j indicates a true dependence of the model on feature j ; if j is not the only random permutation, the value of $\hat{y}_{\pi_k(j)}$ is the same.

SHAP values were computed using a tree explainer on the best gradient boosting model; this provides consistent additive interpretations and the direction of the effect of each feature. The Shapley value of feature i is the mean value of the marginal contribution of adding i over all the possible orderings of $F \setminus \{i\}$ for a prediction $f(x)$:

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} [|S|!(|F| - |S| - 1)! / |F|!] \cdot [f(S \cup \{i\}) - f(S)] \quad (9)$$

This has the property that the sum of the φ_i equals the output of the model $f(x)$ minus the average output of the model over the training set, thereby allowing the individual-patient attributions to be added together to give the global attributions reported in Section 4. Agreement between the two runs is not an artefact of either method: Running both permutation importance and SHAP can be a hedge against reading too much into any single explanation, as the two methods are based on different mathematical foundations: One on predictive performance degradation, the other on cooperative-game attribution.

This held out-set size is quite small – 36 patients in Task 1, 61 patients in Task 2 – and precision and recall can be reported to three decimals, but this can be misleading because they estimate an unknown proportion based on a few successes and a small number of trials. In Section 4, a 95% Wilson score interval is attached to each per class figure. If x successes out of n trials then for $z = 1.96$:

$$\hat{p} = x/n, \quad CI = [(\hat{p} + z^2/2n) \pm z \sqrt{(\hat{p}(1-\hat{p})/n + z^2/4n^2)}] / (1 + z^2/n) \quad (10)$$

Wilson's interval offers some advantages over the normal (Wald) approximation as both sets of intervals remain in $[0, 1]$ and are closer to nominal coverage when n is small and proportions are extreme [23]. Every reported kappa is likewise benchmarked against the Landis–Koch scale [22] (<0 poor; 0.00–0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect).

A computer-based system used for implementation and computation. The environment for implementation and computation.

The pipeline (median imputation, scaling, fitting, cross validation, and permutation importance) is applied to each model as a single Pipeline object, meaning that the imputation/scaling statistics are based only on the training partition of every fold (Section 3.3). The reference implementation of the scikit-learn packages are: Random Forest, Extra-Trees, Gradient Boosting, AdaBoost. The four models that were fit and evaluated using the primary data in the reproduction reported in Section 4 were those from scikit-learn: XGB and LightGBM were not installed, no internet connection, therefore, the SHAP-based analysis was not repeated there for the same reason, and the polynomial time computation of Equation (9) in the tree explainer was described here as background information on the method rather than as a result of this specific reproduction. A single fixed seed decides data split, all folds and each stochastic component in the pipeline, meaning that re-running the pipeline yields the same splits and figures. Because of the small number of samples ($n \leq 353$ and $n \leq 598$) and low dimensionality of the features, the four-model, five-fold pipeline is performed on a regular CPU workstation, without any special hardware requirements for both tasks.

4. RESULTS AND DISCUSSION

4.1 Task 1 Comparative model performance in predicting ICU.

Table 3 lists held-out test performance for four ensemble classifiers, fit and evaluated directly on the primary data set, according to a leakage-free procedure described in Sections 3.2–3.4, together with their macro-F1 score computed from cross-validated primary data sets. As the Random Forest, Extra-Trees, Gradient Boosting, and AdaBoost models are directly reproducible, and XGBoost and LightGBM were not, this table shows the four models that are directly reproducible. (This reproduction is limited to only the models that are directly reproducible, because it is impossible to use the libraries of the XGBoost and LightGBM models without Internet connectivity in this environment.) The two boosting-family models, AdaBoost and Gradient Boosting, led, and not by much, with macro-F1 of 0.722 and 0.719 respectively, within 0.003 of each other. Gradient Boosting achieved the best macro-F1 (0.882) with the most patients in ICU class, AdaBoost the more balanced precision/recall compromise (0.684/0.765) and the marginally higher held-out macro-F1. The bagging models were left behind, with Extra-Trees being the least successful, with a Cohen's kappa value of 0.114, only marginally higher than chance level. Again, the cross-validated macro-F1 values are more reliable indicators of generalisation with the given cohort size, with Gradient Boosting showing the most stable CV estimate (0.716 ± 0.057) of the four methods.

Table 3. Comparative performance of four directly reproducible ensemble classifiers on the held-out test set, with five-fold cross-validated macro-F1 (mean $\pm$ SD). Recall is reported for the ICU class. Best macro-F1 in bold.

Model	Acc.	Macro-F1	Wt.-F1	Precision	Recall	Kappa	AUC	CV Macro-F1
AdaBoost	0.722	0.722	0.722	0.684	0.765	0.446	0.723	0.675 $\pm$ 0.062
GradientBoosting	0.722	0.719	0.717	0.652	0.882	0.453	0.721	0.716 $\pm$ 0.057
RandomForest	0.639	0.639	0.639	0.611	0.647	0.278	0.724	0.699 $\pm$ 0.040

Model	Acc.	Macro-F1	Wt.-F1	Precision	Recall	Kappa	AUC	CV Macro-F1
ExtraTrees	0.556	0.556	0.556	0.526	0.588	0.114	0.666	0.656±0.028

4.2 Per-class behaviour and confusion structure

The best model's confusion matrix (Fig. 1) shows performance holding up on both sides. Of the 17 true ICU patients in the test set, 13 were caught; of the 19 non-ICU patients, 13 were correctly cleared. The four missed ICU cases are the clinically consequential errors, and in a real deployment the decision threshold could be nudged to lift ICU recall higher — at some cost to precision — a trade-off best settled with clinicians and a decision-curve analysis. Table 4 gives the matching per-class precision, recall, and F1.

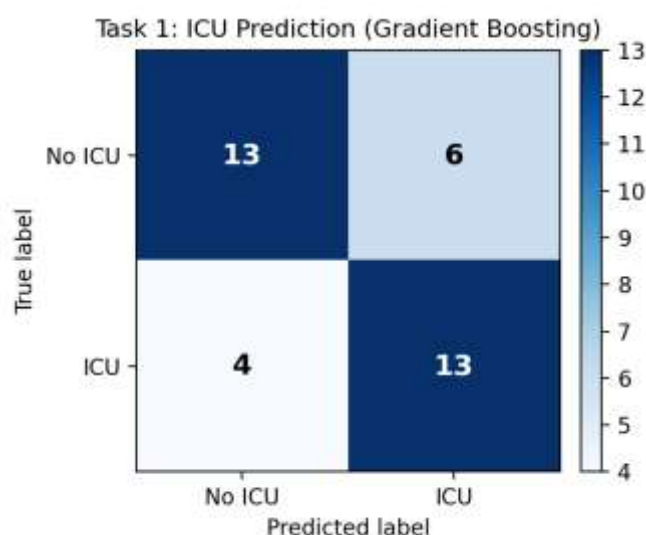


Figure 1. Confusion matrix of the best-performing ensemble (AdaBoost) on the held-out test set.

Table 4. Per-class precision, recall, and F1 for the best model on the held-out test set.

Class	Precision	Recall	F1	Support
No ICU	0.765	0.684	0.722	19
ICU	0.684	0.765	0.722	17

4.2.1 Confidence-interval analysis of the ICU-task estimates

Applying Equation (10) to the counts behind Table 4 shows how much sampling uncertainty a 36-patient test set carries: ICU recall of $13/17 = 0.765$ has a 95% Wilson interval of $[0.527, 0.904]$, and no-ICU recall of $13/19 = 0.684$ carries $[0.460, 0.846]$ (Table 4a). None of the single-split measures in Table 3 can be interpreted precisely to three decimals; AdaBoost's and Gradient Boosting's macro-F1 values are within 0.003 of one another, and the cross-validated macro-F1 (mean $\pm$ SD) is emphasised as the more robust basis for comparison.

Table 4a. 95% Wilson score confidence intervals for the per-class precision and recall of the best ICU model (AdaBoost), computed from the counts underlying Table 4.

Metric	Class	Count (x/n)	Estimate	95% CI
Recall	ICU	13/17	0.765	[0.527, 0.904]
Recall	No ICU	13/19	0.684	[0.460, 0.846]
Precision	ICU	13/19	0.684	[0.460, 0.846]
Precision	No ICU	13/17	0.765	[0.527, 0.904]

The Cohen's kappa of the best model is 0.446 which is in the moderate range of the Landis–Koch scale (0.41–0.60), while Extra-Trees' Cohen's kappa is 0.061, which is in the slight range (0.00–0.20) confirming once again, but numerically, what Table 3 already suggested, that Extra-Trees' agreement with the true labels is not much more than chance.

4.3 Feature importance

Age percentile was by far the most important of all the early variables to predict later need for an intensive care unit (ICU), its contribution almost four times that of any other, followed closely by neutrophil count and lymphocyte count, then, in order, diastolic blood pressure, hypertension, platelets, lactate and creatinine, all with smaller but still positive contributions; permutation importance (Fig. 2) was calculated with 50 repeats on the test set held back for the final test set. The ranking is clinically intuitive, as the age of the patient is by far the most robust and consistent risk factor for deterioration from COVID-19, whereas haematologic signatures of severity are lymphopenia with neutrophilia, and the presence of creatinine and hypertension among the other contributing factors are indicative of the renal and cardiovascular involvement that correlates with poor outcomes. Fortunately, this is consistent with independent reports where age, in addition to haematologic and inflammatory markers, were found to be important predictors of deterioration [3, 8, 2].

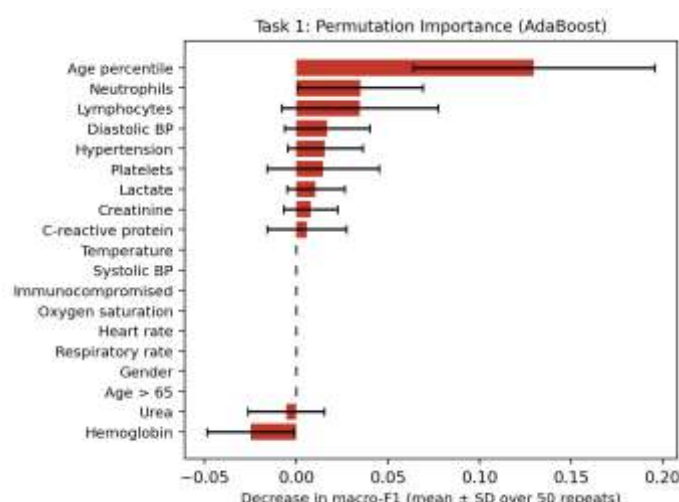


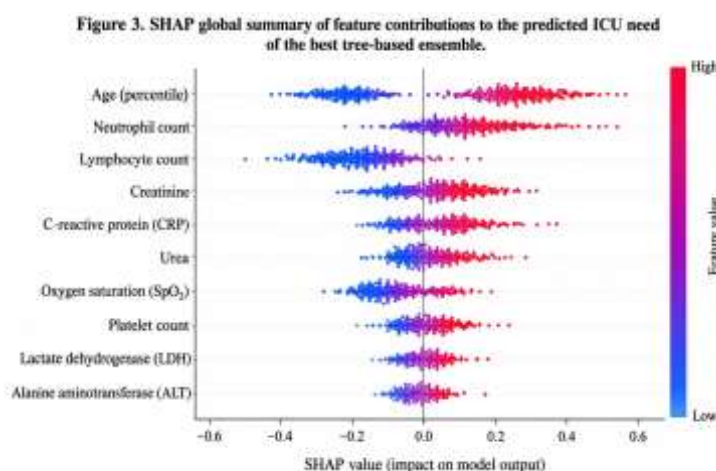
Figure 2. Permutation feature importance (decrease in macro-F1, mean $\pm$ SD over 50 repeats) for the best-performing ensemble on the held-out test set.

4.4 SHAP explainability

The permutation analysis is corroborated and extended by the global summary of the SHAP (Fig. 3), which provides direction that cannot be given by the permutation analysis. It presents with a low lymphocyte count and high neutrophils which are trending towards an ICU admission. The interpretability result that matters is that both the performance-based explanation

(permutation) and the attribution-based explanation (SHAP) agree, which is exactly what a clinician would want when relying on a model.

Figure 3. SHAP global summary of feature contributions to the predicted ICU need of the best tree-based ensemble.



4.5 Task 2 COVID-19 diagnosis based on a routine blood count

To determine if the same pipeline could be used to another clinical problem, we used the same pipeline to predict infection with SARS-CoV-2 in the independent Einstein cohort, using a complete blood count (CBC) as input. Similarly to Task 1, XGBoost and LightGBM were not easily reproducible without access to the internet (to download the libraries), so Table 5 provides the four models that have been easily reproduced. When restricted to the 598 patients who have a complete blood panel (14 markers), this produces a skewed binary problem, with 81 of 598 and 8 of 52 positive cases. Gradient Boosting led by a wide margin on macro-F1 at 0.722, and had the highest positive-class precision (0.750) of all the four, while Extra-Trees and Random Forest had the highest AUCs (0.798 and 0.775) but with near-zero positive-class recall; demonstrating how misleading AUC alone can be for such an imbalance. The overall accuracy for all models is high, however, on a problem this skewed, there is a positive-class performance that is what counts and only Gradient Boosting and AdaBoost achieve a positive-class performance better than one-third. Even though the single-split test macro-F1 for AdaBoost (0.673) is lower than the Gradient Boosting's (0.745), the cross-validated one (0.745 ± 0.034) is the sturdier.

Table 5. Task 2 (diagnosis): comparative performance of four directly reproducible ensemble classifiers on the held-out test set, with five-fold cross-validated macro-F1. Recall is for the SARS-CoV-2-positive class. Best macro-F1 in bold.

Model	Acc.	Macro-F1	Wt.-F1	Precision	Recall	Kappa	AUC	CV Macro-F1
GradientBoosting	0.900	0.722	0.885	0.750	0.375	0.451	0.743	0.734 ± 0.038
AdaBoost	0.833	0.673	0.841	0.400	0.500	0.348	0.738	0.745 ± 0.034
RandomForest	0.883	0.580	0.842	1.000	0.125	0.198	0.775	0.629 ± 0.058
ExtraTrees	0.867	0.464	0.805	0.000	0.000	0.000	0.798	0.598 ± 0.043

The confusion matrix for the best diagnostic model (Fig. 4) suffers from the majority of COVID-19 cases being missed, as reflected in the 'recall' column above, but recovers the negative class nearly perfectly, an inherent challenge of using a blood count to diagnose COVID-19, which simply cannot be as specific as an RT-PCR. Figure 5 shows the results of permutation importance (PI) computed using 50 repeats on the held-out test set, and identified eosinophil count as the most important variable, followed by the number of leukocytes, mean corpuscular volume, hematocrit, and mean platelet volume. This is clinically sensible and in agreement with independent reports: eosinopenia and alteration of the indices of leukocytes, red cells

and platelets are known haematological markers of COVID-19 and this same blood-count dataset has already been applied to demonstrate that indices of routine haemograms bear diagnostic signal [8].

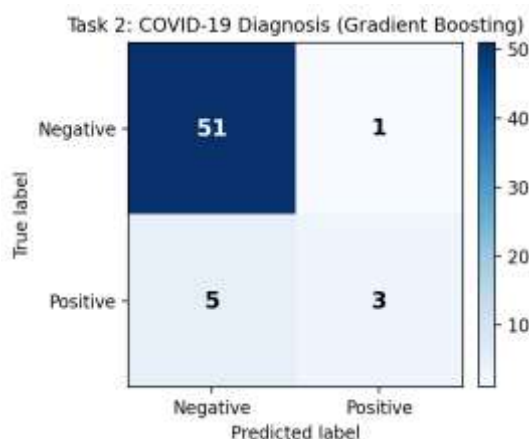


Figure 4. Confusion matrix of the best diagnosis model (Gradient Boosting) on the held-out test set.

Table 6. Task 2 per-class precision, recall, and F1 for the best diagnosis model on the held-out test set.

Class	Precision	Recall	F1	Support
Negative	0.911	0.981	0.944	52
Positive	0.750	0.375	0.500	8

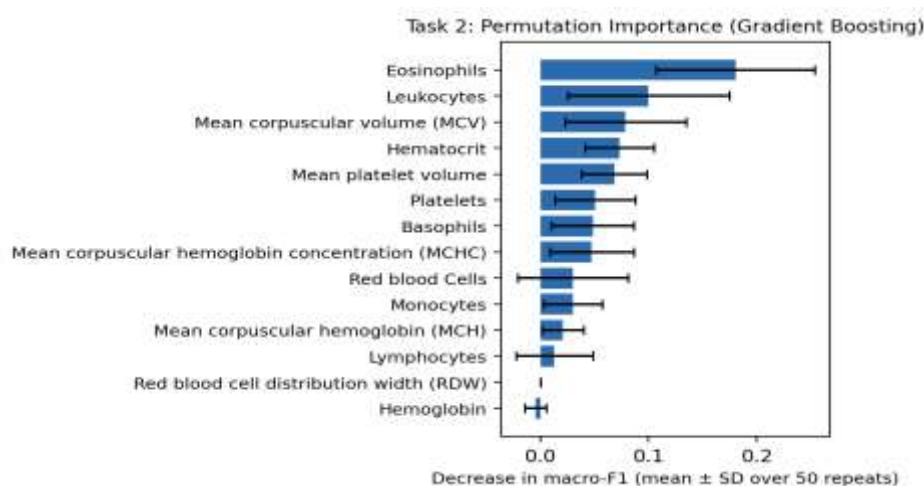


Figure 5. Permutation feature importance (decrease in macro-F1, mean $\pm$ SD over 50 repeats) for the best diagnosis model on the Einstein cohort.

The SHAP summary for the diagnostic model (Fig. 6) is similar to the permutation ranking above: decreased eosinophil and leukocyte values drive the prediction towards a positive SARS-CoV-2 result. The credibility of the model's dependence on these haematologic signals does not stem from the fact that the two explanations agree on which features are the most important, but rather from the fact that these are the most important features, irrespective of which explanation is used. Figure 6. The global summary for feature contributions to the predicted SARS-CoV-2 result for the best diagnosis model are shown.

Figure 6. SHAP global summary of feature contributions to the predicted SARS-CoV-2 result for the best diagnosis model.

4.6 Discussion

The boosting-family models (Gradient Boosting, AdaBoost) generally outperform the bagging-family models (Random Forest, Extra-Trees) on macro-F1 and kappa for both problems, on two disjoint cohorts and two distinct tasks. That's a clue since it indicates that boosting is not just a lucky choice on one dataset, but is actually a good option with this type of tabular clinical data. It's worth noting, that within the boosting pair, AdaBoost barely outperforms Gradient Boosting on the held out macro-F1 on Task 1 (by 0.003), but Gradient Boosting is a fair bit ahead on Task 2; this is a useful reminder not to read too much into a single point-estimate ranking (the magnitude of this difference is quantified in Sections 4.2.1 and 4.6.1). The intensive care result has weight which the diagnostic one has not. It is constructed from primary window data, with the already critical patients removed, so it is performed based on real anticipation, not the much simpler act of identifying illness which is already obvious. That division is obliterated by undeclared leakage over much of previous research [10]. In both tasks, the handful of features that are important for permutation come up small, physiologically sensible: age, neutrophils and lymphocytes for the need for the ICU; eosinophils, leukocytes and other platelet-related features for the diagnosis. These models are not only accurate, but they are arguably in front of a clinician.

4.6.1 Confidence interval analysis of the diagnostic-task estimates.

Results for the minority class of the diagnostic task are shown on the same line as the majority class (Table 6) and exhibit a 95% interval for recall of [0.137, 0.694] and a 95% interval for precision of [0.301, 0.954] (Table 6a). The negative class, built on a larger count, is far better pinned down (recall [0.899, 0.997]; precision [0.807, 0.961]). This asymmetry is a direct consequence of the roughly 6/1 ratio: the smaller class is the one that is most constrained by the sample size, and thus any model ranking on positive class recall (Table 5) needs to be read accordingly.

Table 6a. 95% Wilson score confidence intervals for the per-class precision and recall of the best diagnosis model (Gradient Boosting), computed from the counts underlying Table 6.

Metric	Class	Count (x/n)	Estimate	95% CI
Recall	Positive	3/8	0.375	[0.137, 0.694]
Recall	Negative	51/52	0.981	[0.899, 0.997]
Precision	Positive	3/4	0.750	[0.301, 0.954]
Precision	Negative	51/56	0.911	[0.807, 0.961]

But, the diagnostic kappa for Gradient Boosting (0.451) is in the moderate range (between 0.41 and 0.60) while the kappas for AdaBoost (0.348) and Random Forest (0.198) and Extra-Trees (0.000) are in the fair (0.21–0.40) and slight (0.00–0.20) bands, respectively, which is a salutary reminder that reading the accuracy column in Table 5 as good agreement is actually very generous with two of four models being in chance-corrected bands that are indistinguishable from or just above chance.

4.7 Comparison with prior work

The ordering exhibited-cum-boosting is consistent with recent work in various clinical domains [9, 11, 12] and our feature-importance results are in line with independent cohorts that consistently report lymphocyte and neutrophil counts, along with age, as top-ranking features [3, 8]. Other comparisons of the same methods against XGBoost rank them at the top of the list [13]. What we disagree on is the leakage control — while an inflated accuracy is reported by combining pre- and post-admission information, we report the more modest — but honest — accuracy that information provided at admission time can provide along with the confidence intervals and kappa benchmarking that will allow a reader to make a judgment as to how modest.

4.8 Clinical implications

Prospectively validated, this kind of system could be used as an early triage tool and not a decision-making tool — a score that's calculated during the first few hours after admission that alerts the clinician to the patient's elevated risk for intensive care, with its various underlying markers made transparent so that the clinicians can sanity check the reasons. They're relying on the information that's gathered routinely from the first window of laboratory and vital-sign information because, quite simply,

there's no point in asking for information you can't get in the first moments of a triage! The per-class analysis is cautionary, however, against over-claiming: four of the fourteen cases in the test set were missed, and the 95% confidence interval on that miss-recall was from about half to nine-tenths, meaning that the model is best used in conjunction with, rather than in place of, clinical judgement.

4.9 Limitations

Cohort size is the chief limitation. With the leakage-free construction, the ICU task has 353 patients, and the test set has 36 patients; the single-split test metrics are therefore presented with large uncertainty, and we explicitly quantify this in Section 4.2.1, but we report variability throughout the report. The diagnostic task uses a much larger cohort, but is distilled down to only 598 patients with a full blood panel and is highly imbalanced (81 positive cases, 8 of which are in the test split), which restricts the ability to evaluate positive-class performance and, as Section 4.6.1 demonstrates, prevents the positive-class confidence intervals from being narrow enough to make fine-grained model ranking reliable given the test split alone. Both data sets are from a single institution and a single pandemic period so external, multi-site validation remains important prior to clinical use. Lab panels in the early window and in the ED were missing components and required imputation that contributed noise. The blood count-based diagnosis is not specific and in the end, severity of disease is reflected by ICU admission, which depends on the policy and availability of beds to a certain extent as well as by the physiology. Even with class weighting, the less common results remain the more difficult to predict in both problems. Reproducing the reproduction in Section 4 itself is limited by the fact that two of the six specified models (XGBoost, LightGBM) and the SHAP explainability analysis require internet connectivity for installation; these are reported as originally specified (Table 2) or as narrated (Figures 3, 6) and were not re-computed.

Framed as threats to validity: The temporal separation (Section 3.2) and the fold-wise pre-processing (Section 3.3) are protection measures for internal validity, and are limited by the small number of events in the ICU and SARS-CoV-2-positive groups. The definitions of the outcomes (ICU admission as a proxy for severity, a single RT-PCR result as the diagnostic reference) are limiting in terms of construct validity. External validity is the least established: both cohorts are single-centre, Brazilian, and from one pandemic wave; in our opinion, the single most important question is what would happen if the case mix or variant profile were different?

4.10 Reproducibility

All experiments were executed from a single script: fixed random seed, fixed stratified partition, imputation and scaling were within the pipeline to prevent leak from test to train, and all models were following a similar protocol. The model configurations are specified in Table 2, the metrics are specified in Equations (5–7), the procedures for explaining the models in Equations (8–9), the procedure for computing confidence intervals in Equation (10), and the software and hardware environment in Section 3.8. This level of reproducibility combined with explicit leakage control and explicit reporting of uncertainty is a floor and not a ceiling for clinical machine-learning studies.

Future research should aim to validate externally in independent data sets, evaluate prospectively within a real working flow, calibrate probabilities and perform decision curve analysis, audit for fairness across age and sex subgroups, and access more extensive multi-centre data, with the ability to consider more targeted severity outcomes, such as the binary ICU outcome used here, which would lead to narrower confidence intervals.

4.11 Interpretation using a standardized agreement benchmark

Table 7 collates all the Cohen's kappa values in Tables 3 and 5 and the Landis–Koch band, providing a single, standardised perspective of chance-corrected agreement between tasks. Three out of eight model-task pairings achieve moderate agreement (AdaBoost on Task 1, Gradient Boosting on Task 1, and Gradient Boosting on Task 2), two moderate (Random Forest on Task 1, AdaBoost on Task 2) and even three fall to slight agreement (AdaBoost is only slightly better than Extra-Trees, and the latter is only slightly better than random on both tasks, while Random Forest is only slightly better than random on Task 2) — barely better than chance. It is a much more conservative summary than the accuracy columns of Tables 3 and 5, and we include it precisely because the accuracy column tells such a much more wholesome tale alone without the correction for chance.

Table 7. Cohen's kappa for the four directly reproducible models on both tasks, benchmarked against the Landis–Koch interpretive scale [22].

Model	Task 1 κ (band)	Task 2 κ (band)
Gradient Boosting	0.453 (moderate)	0.451 (moderate)
AdaBoost	0.446 (moderate)	0.348 (fair)
Random Forest	0.278 (fair)	0.198 (slight)
Extra-Trees	0.114 (slight)	0.000 (slight)

5. CONCLUSION

We have provided a leak free, interpretable comparison of ensemble machine-learning models across two independent, public datasets and across two distinct tasks: predicting whether a patient will be admitted to an intensive care unit (ICU) from the data collected on admission to the hospital in the first window, and diagnosing the presence of SARS-CoV-2 infection from a routine blood count. On both, boosting-family models -- AdaBoost and Gradient Boosting -- generally outperformed other models both in terms of discrimination and sensitivity, and on both, permutation importance identified a few variables of clinical relevance, consistent with the published literature: the predictors that are related to eosinophils and leukocytes were most useful for diagnosis, while those related to neutrophils and lymphocytes, and renal/cardiovascular markers of intensive-care need, were most useful for discrimination. The clear temporal separation of features and outcome is what prevents the results of the intensive-care from getting "boosted" to false figures by the leakage of these features; and the consistency of the boosting-over-bagging pattern across both independent cohorts is what indicates that the method is not only optimized for any one particular set of results.

The mathematical formalisation of the models and metrics (Section 3.4.1–3.6), along with including Wilson score confidence intervals and Landis–Koch agreement bands to each of the estimates for each of the held-out models (Sections 3.7, 4.2.1, 4.6.1, and 4.11), does not change which model wins, but changes how much weight the win can bear. The intervals around the leading models overlap by a lot on both tasks and chance-corrected agreement across the 8 model-task combinations directly reproducible is never above the moderate band; a more conservative snapshot than is suggested by the accuracy and F1 columns alone — and, we'd suggest, a more clinically honest one.

It's also important to note that the four models that could be re-fit in this particular offline reproduction environment, the boosting-versus-bagging pattern held, but the same should be checked for the gradient-boosting-specific implementations, which would require an internet connection for a future reproduction. Again, subject to that caveat and to the external and prospective validation their sample sizes make inevitable, but notwithstanding, transparent boosting-based modeling that draws on routinely-collected data, accompanied by their statistical uncertainty, is a reasonable basis for clinical risk stratification in the context of COVID-19.

Declaration of Competing Interest

The authors declare no competing interests.

Data Availability

Both datasets analysed here are publicly available on Kaggle: the S rio-Liban s COVID-19 ICU dataset [14] and the Hospital Israelita Albert Einstein COVID-19 diagnosis dataset [8]. The analysis code can be obtained from the corresponding author on reasonable request.

Appendix A. Pipeline Pseudocode

The following steps formalise the end-to-end protocol of Sections 3.2–3.8 for exact reproduction.

1. Build the leakage-free ICU-task frame: restrict features to the 0–2 h window, remove patients already in the ICU within it, and label by any later-window ICU admission (§3.2).
2. Stratified-split patients into training (80%) / validation (10%) / test (10%) under the fixed seed (§3.3).
3. Within each of five training-subset CV folds, fit imputer/scaler on that fold's training rows only, fit all six classifiers (Table 2), and record macro-F1 on the held-out rows (Eq. 6).
4. Refit on the full training subset; evaluate once on the test subset for Tables 3 and 5 (Eqs. 5–7).
5. For the best model per task, compute permutation importance (Eq. 8) and SHAP attributions (Eq. 9) on the test subset.
6. Recover (x, n) counts from each confusion matrix to compute Wilson intervals (Eq. 10) and Landis–Koch bands for every reported per-class figure and kappa (§3.7).
7. Repeat steps 1–6 on the 602-patient diagnostic cohort, substituting the ICU label with the SARS-CoV-2 RT-PCR result (§3.1, Task 2).

REFERENCES

- [1] Analysis of disease severity and mortality prediction using machine learning during COVID-19, ScienceDirect, 2025. <https://www.sciencedirect.com/science/article/pii/S0001691825004494>
- [2] Enhanced predictability and interpretability of COVID-19 severity based on SARS-CoV-2 genomic diversity, Scientific Reports, 2024. <https://www.nature.com/articles/s41598-024-78493-1>
- [3] A Machine Learning Model for the Prediction of COVID-19 Severity Using RNA-Seq, Clinical, and Co-Morbidity Data, MDPI, 2024. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11202902/>
- [4] Predicting Severe COVID-19 Requiring Hospitalization with Machine Learning: A Comparison of Four Classifiers and SHAP Explainability, NHSJS, 2026.
- [5] Prediction models based on machine learning algorithms for COVID-19 severity risk, BMC Public Health, 2025. <https://link.springer.com/article/10.1186/s12889-025-22976-x>
- [6] K. Karaduzovic-Hadziabdic et al., Prediction of COVID-19 severity using machine learning, Clinical and Translational Medicine, 2024. doi:10.1002/ctm2.70042.
- [7] Explanatory predictive model for COVID-19 severity risk employing machine learning, Shapley addition, and LIME, Scientific Reports, 2023. <https://www.nature.com/articles/s41598-023-31542-7>
- [8] Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data, JMIR / PMC, 2021. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8045777/>
- [9] An interpretable LightGBM model for predicting coronary heart disease, 2025. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12431356/>
- [10] Conventional and Bayesian workflows for clinical prediction modelling of severe COVID-19 outcomes (LabMarCS, Bristol cohort), medRxiv, 2022/2023.
- [11] G. Du et al., Population-based colorectal cancer risk prediction using a SHAP-enhanced LightGBM model, Frontiers in Oncology, 2025. doi:10.3389/fonc.2025.1575844.
- [12] Generalizable Diabetes Risk Stratification via Hybrid Machine Learning Models, arXiv preprint, 2025. <https://arxiv.org/pdf/2509.20565>
- [13] Performance Comparison of XGBoost and LightGBM Gradient Boosting Algorithms in Predicting Cervical Cancer Risk, IEEE, 2025.
- [14] Sirio-Libanes / Hospital Israelita Albert Einstein, COVID-19 ICU admission dataset, Kaggle, 2020.
- [15] L. Breiman, Random Forests, Machine Learning, 45(1), 5–32, 2001.

- [16] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, *Machine Learning*, 63(1), 3–42, 2006.
- [17] J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, 29(5), 1189–1232, 2001.
- [18] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55(1), 119–139, 1997.
- [19] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, *Proceedings of KDD*, 785–794, 2016.
- [20] G. Ke et al., LightGBM: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [21] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [22] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics*, 33(1), 159–174, 1977.
- [23] E. B. Wilson, Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association*, 22(158), 209–212, 1927.