Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests

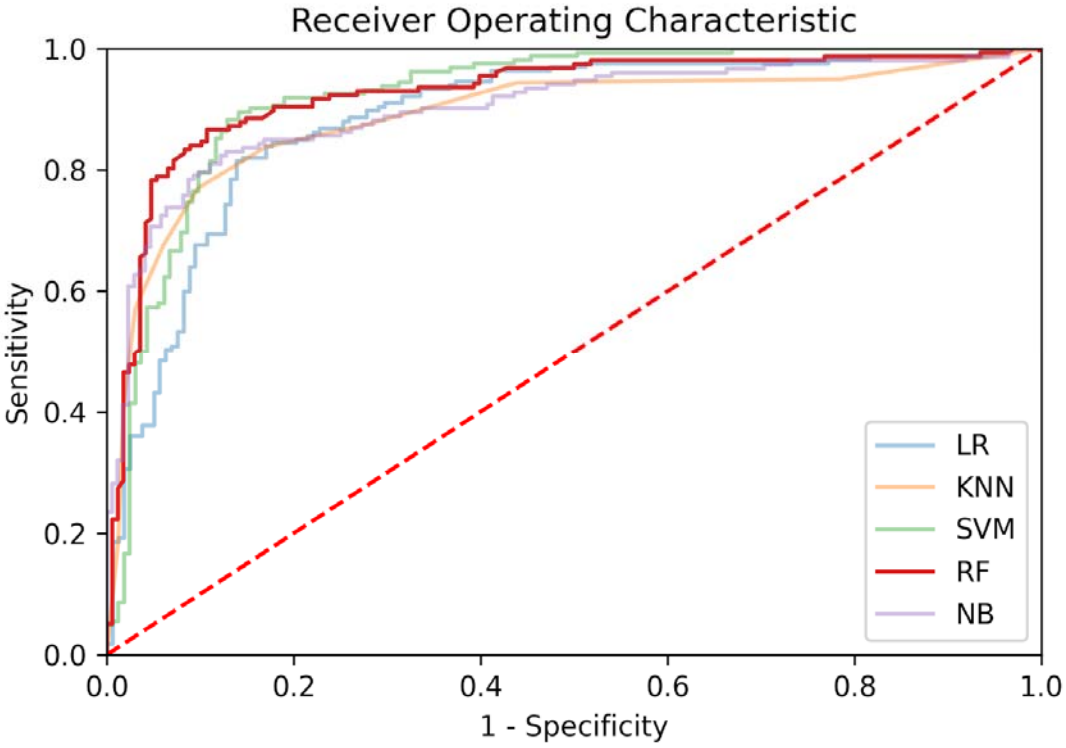
Cabitza Federico[1], Campagner Andrea[2], Ferrari Davide[3], Di Resta Chiara[4], Ceriotti Daniele[5], Sabetta Eleonora[5], Colombini Alessandra[2], De Vecchi Elena[2], Banfi Giuseppe[2], Locatelli Massimo[5], Carobene Anna[5]

SUPPLEMENTAL MATERIAL

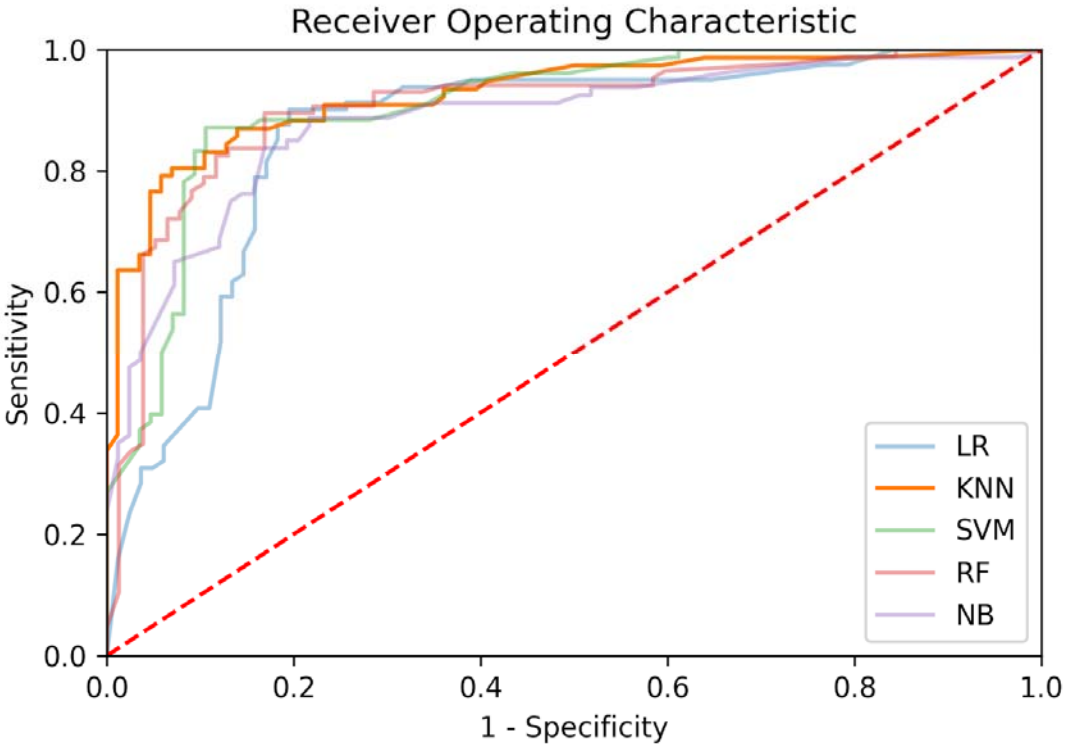
Suppl. Table 1. Hyper-parameters of the machine learning models under consideration.

Model	Random Forest	Naive Bayes	Support Vector Machine	Logistic Regression	k-Nearest Neighbor
Hyper-parameters	Number of estimators, Maximum tree depth, Split criterion, Maximum number of features	/	Kernel, Maximum polynomial degree, Kernel coefficient, Regularization parameter	Regularization penalty, Regularization parameter	Distance weighting, Nearest Neighbor algorithm, Number of neighbors

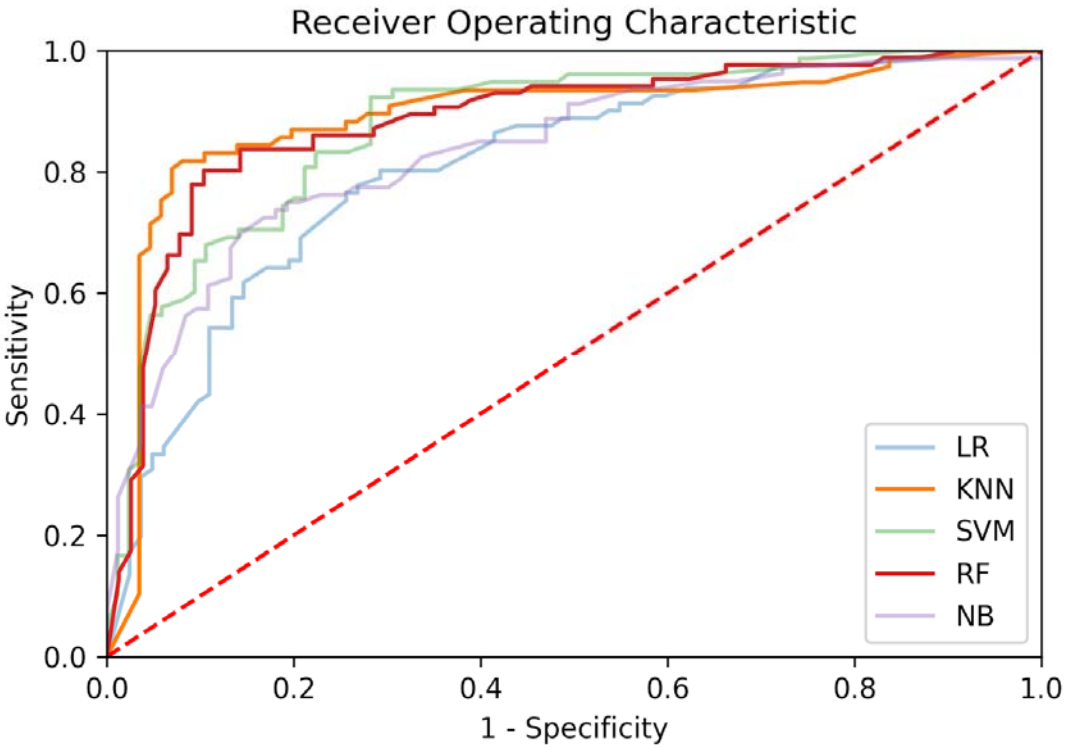
Supplementary Figures



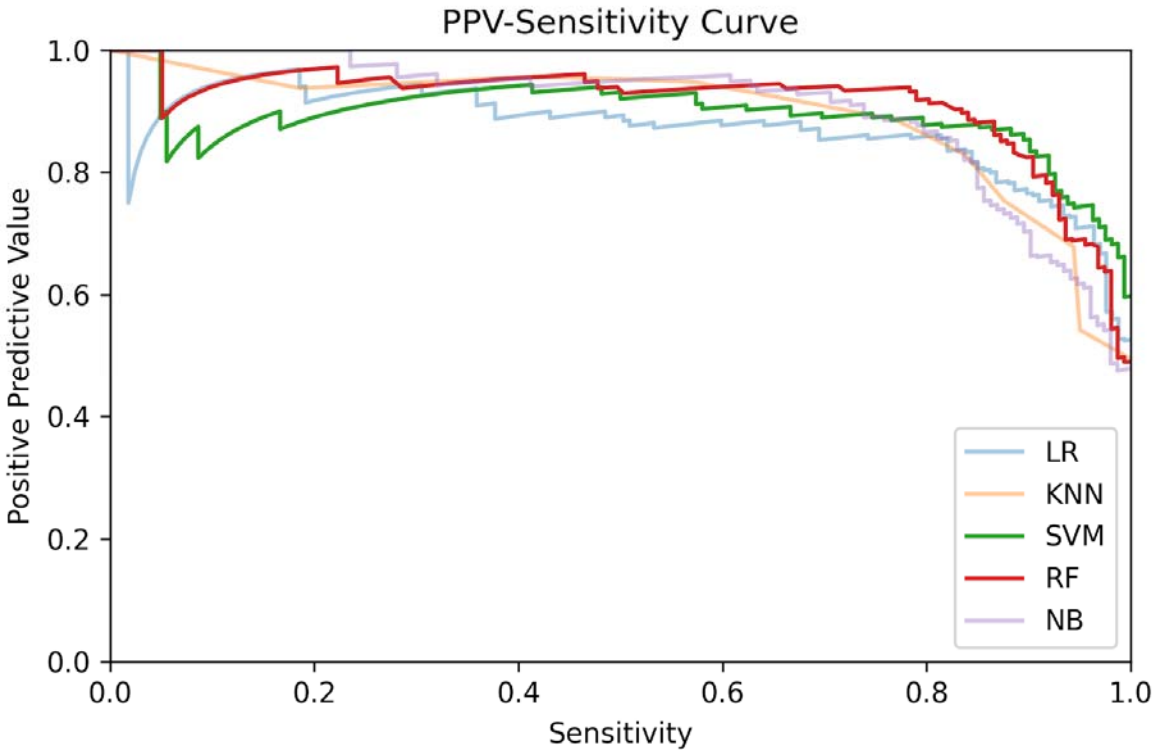
Suppl. Figure 1. Receiver operating characteristic curves for the models trained using the *OSR dataset*.



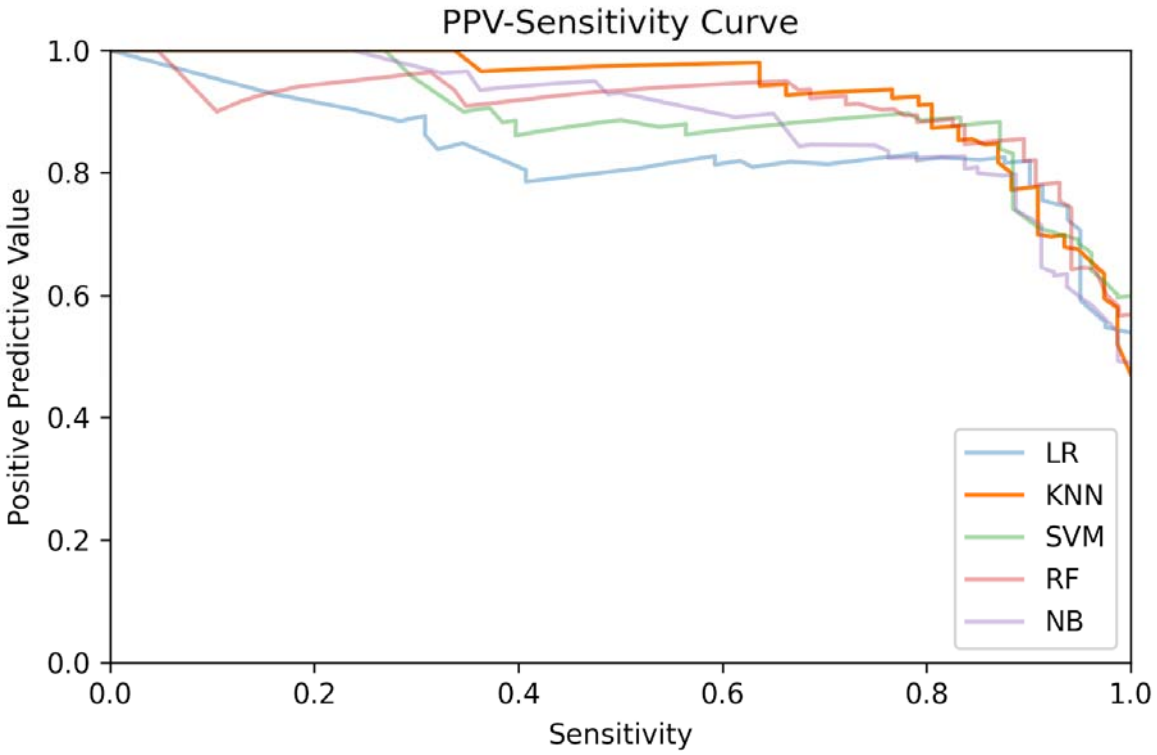
Suppl. Figure 2. Receiver operating characteristic curves for the models trained using the *COVID-specific dataset*.



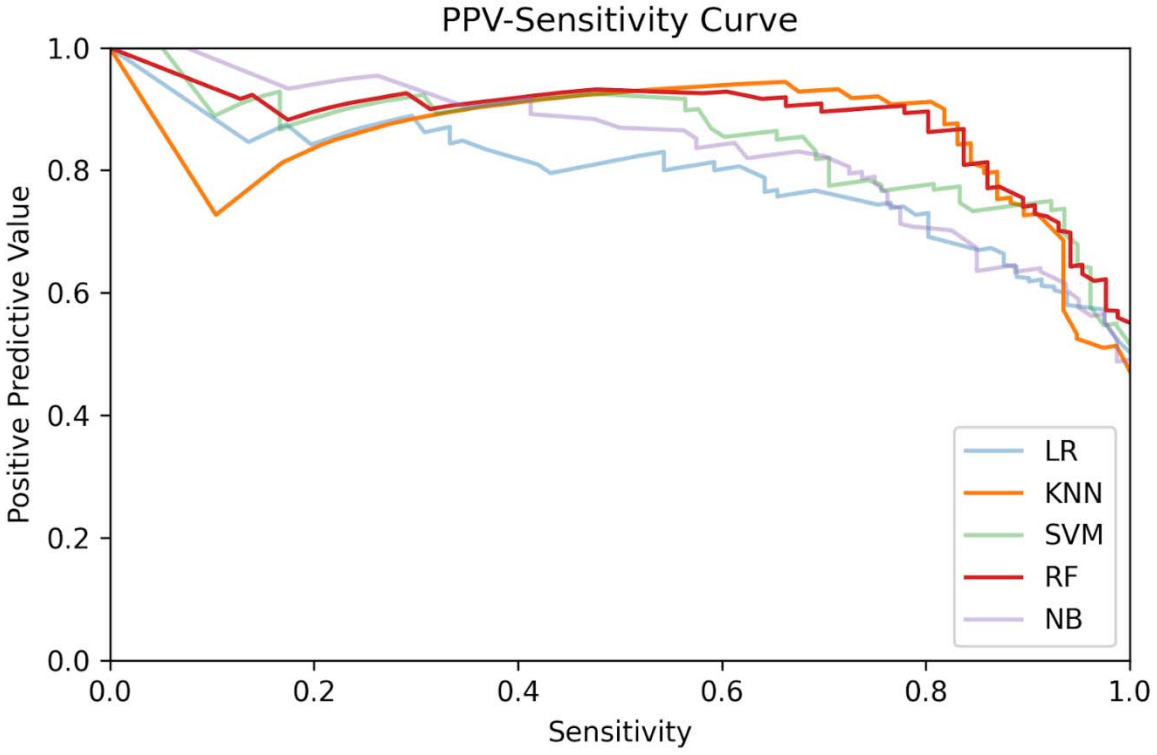
Suppl. Figure 3. Receiver operating characteristic curves for the models trained using the *CBC dataset*.



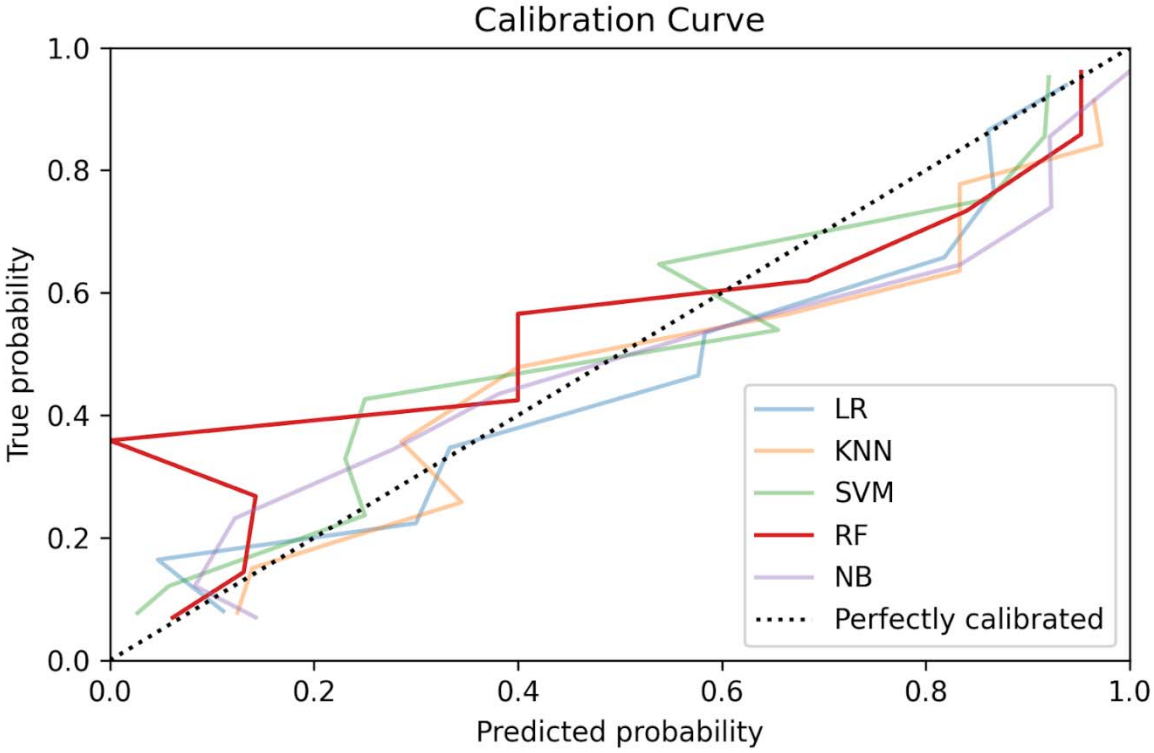
Suppl. Figure 4. Positive predictive value-sensitivity curves for the models trained using the *OSR dataset*.



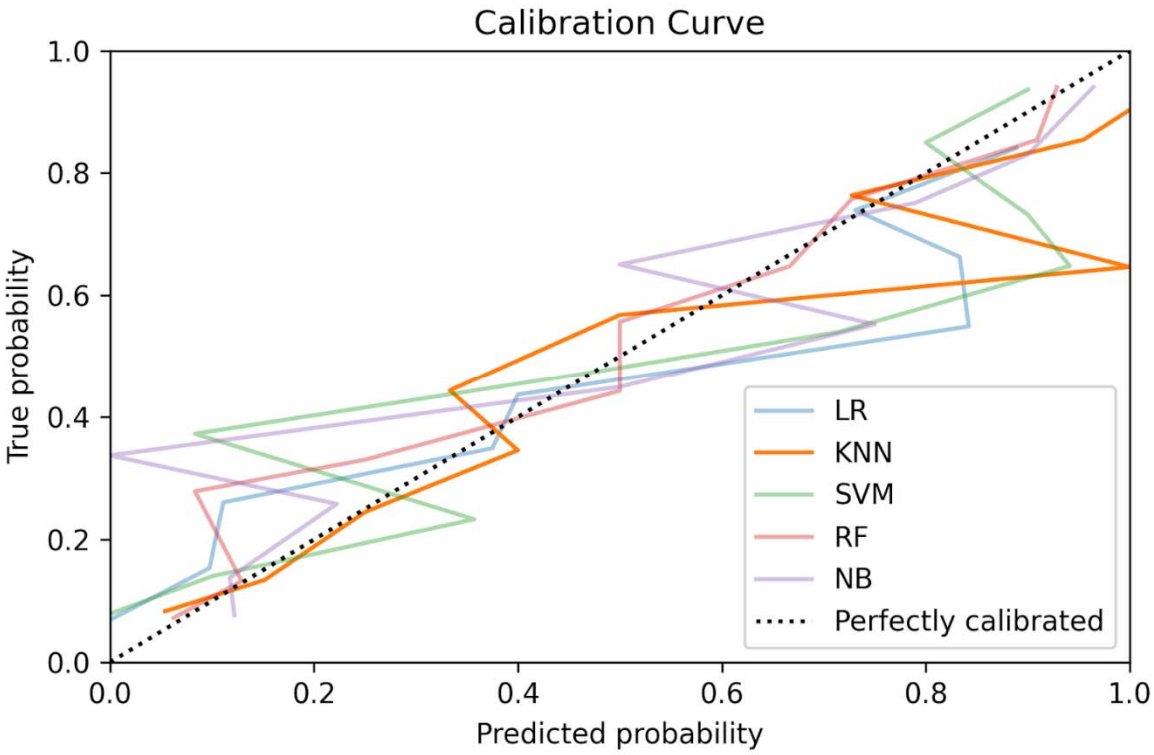
Suppl. Figure 5. Positive predictive value-sensitivity curves for the models trained using the *COVID-specific dataset*.



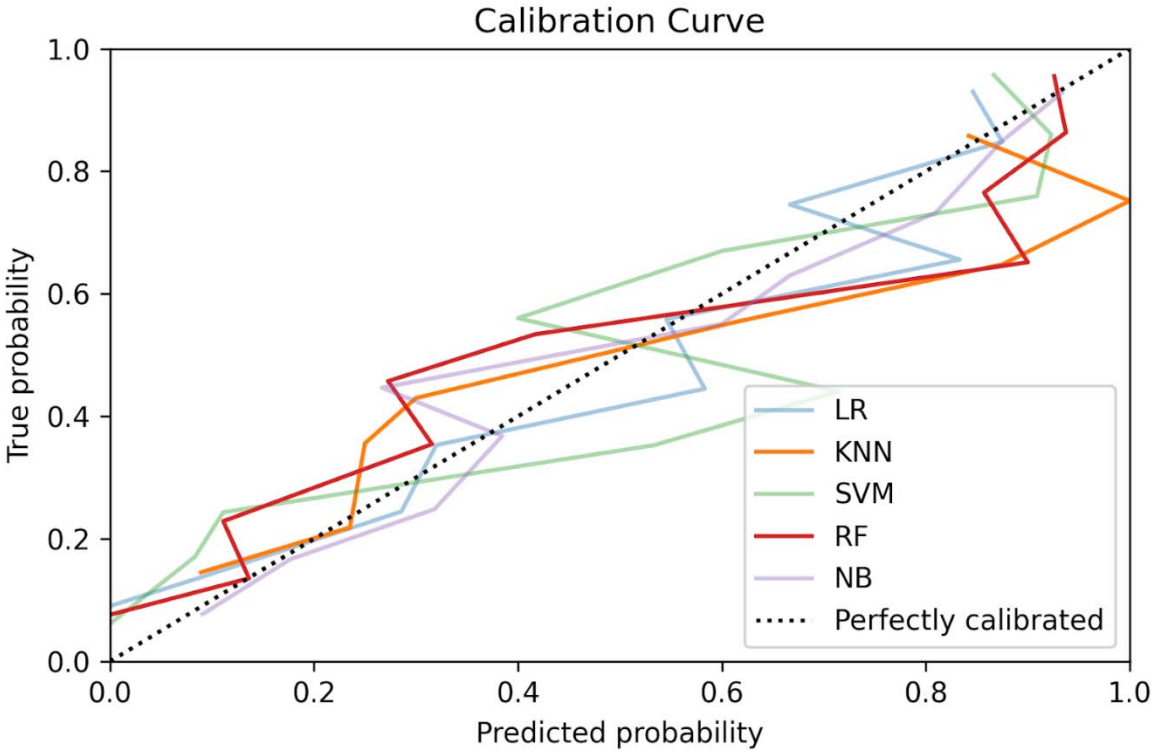
Suppl. Figure 6. Positive predictive value-sensitivity curves for the models trained using the *CBC dataset*.



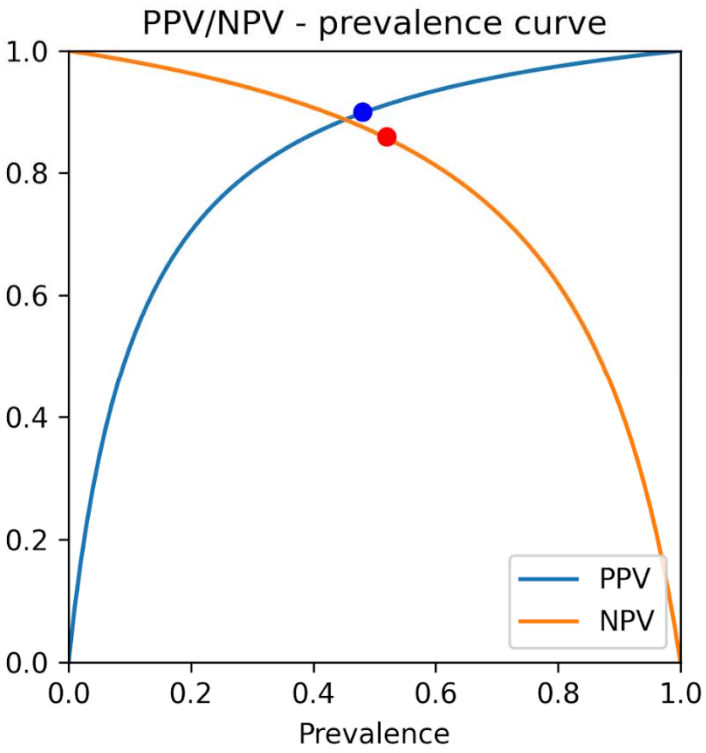
Suppl. Figure 7. Calibration curves for the models trained using the *OSR dataset*.



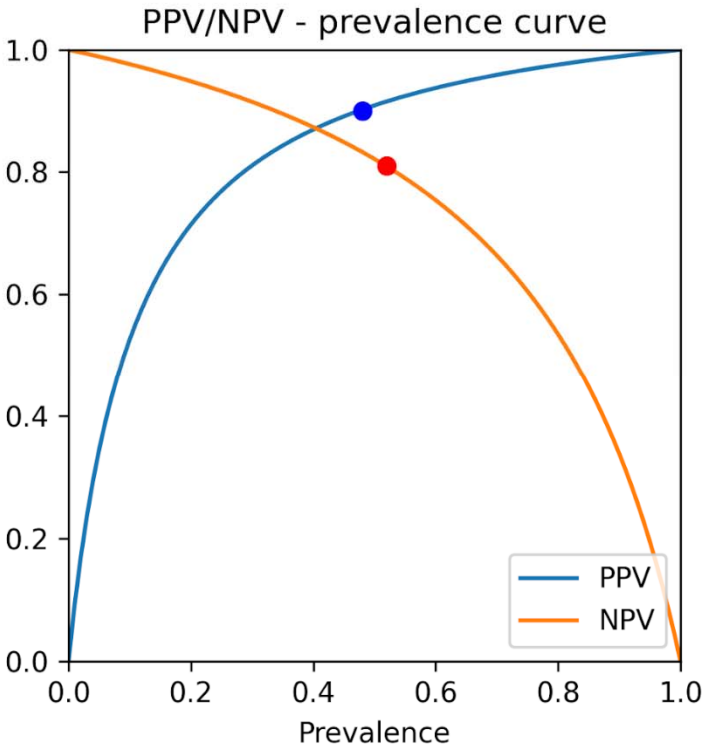
Suppl. Figure 8. Calibration curves for the models trained using the *COVID-specific dataset*.



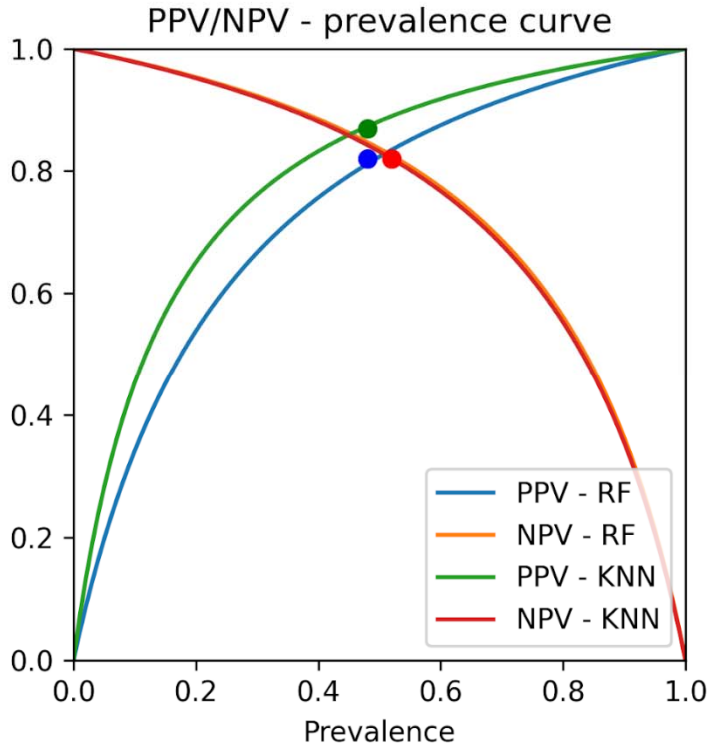
Suppl. Figure 9. Calibration curves for the models trained using the *CBC dataset*.



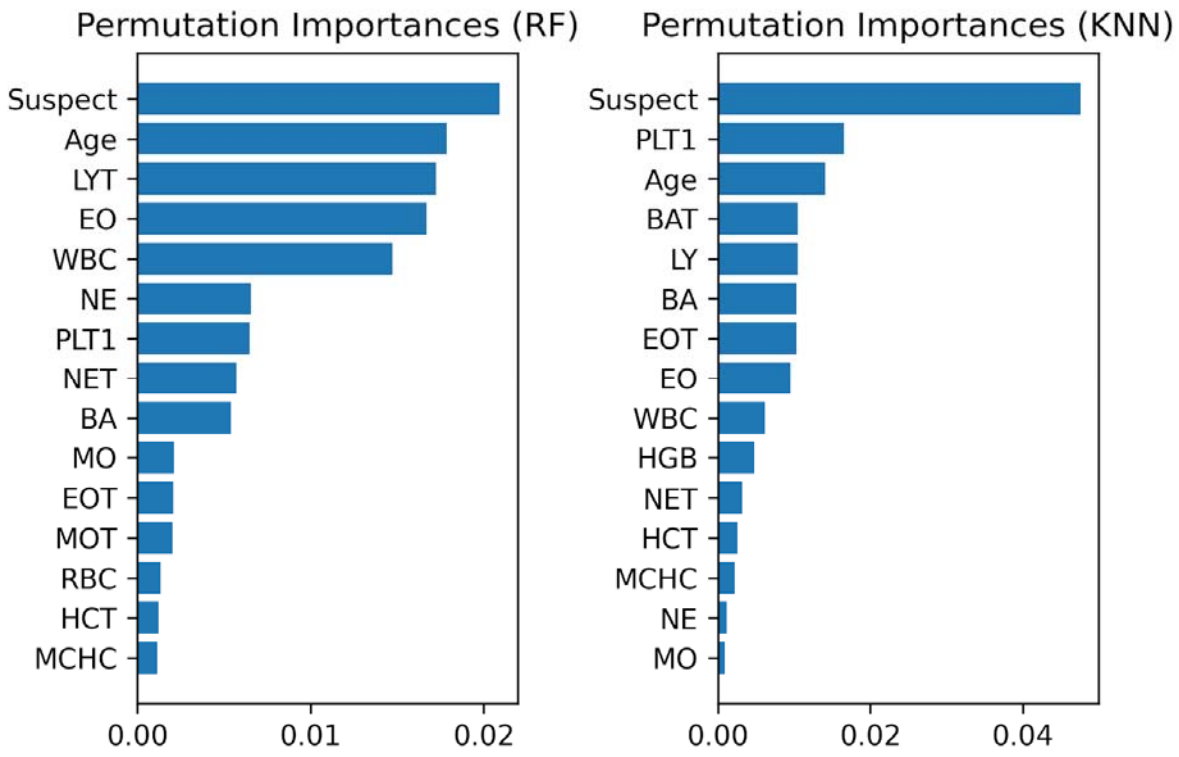
Suppl. Figure 10. Positive predictive value/negative predictive value-prevalence curve for the random forest algorithm, trained using the *OSR dataset*. The points on the curves indicate the prevalence in the dataset.



Suppl. Figure 11. Positive predictive value/negative predictive value-prevalence curve for the k-nearest neighbors' algorithm, trained using the *COVID-specific dataset*. The points on the curves indicate the prevalence in the dataset.



Suppl. Figure 12. Positive predictive value/negative predictive value-prevalence curve for the random forest and k-nearest neighbors' algorithms, trained using the *CBC dataset*. The points on the curves indicate the prevalence in the dataset.



Suppl. Figure 13. Feature importances for Random Forest and k-Nearest Neighbors, trained on the *CBC dataset*.

Identification of the Uncertain Cases

In the ground truthing process, we identified 165 uncertain cases for which we combined the results of the rRT-PCR test together with the radiologic gold standard. The uncertain cases were identified through two different methods: either patients who turned out to be positive within 72 hours after a first negative test and were admitted as inpatients despite this test result; or patients who, despite having had a negative test, had a hematochemical profile that was more similar to positive patients. For this purpose we used the k-means clustering algorithm (k = 2) based on a set of COVID-19 characteristic biomarkers (AST, lymphocytes, calcium, LDH, CRP, WBC, XDP, fibrinogen) [20,21].

Implementation of the Internal-External Validation

The internal-external validation was performed based on the *IOG dataset*, using a bootstrap-based procedure. The goal of this procedure was to evaluate the ability of the developed models to generalize to new settings when provided with a limited amount of new data.

First, we generated 100 random, 50/50 train-test splits of the *IOG dataset*, then for each of these splits: first, the train set of the *IOG dataset* was oversampled using the SMOTE algorithm to obtain a sample of 1,624 synthetic instances; second, the oversampled train set was combined with the *COVID-specific* (respectively, *CBC*) dataset to obtain a combined training set encompassing 3,248 instances; third, the best models (obtained as described in the Methods and Results sections) were re-trained over the combined training set and evaluated on the test set. The average results over the 100 generated splits were reported.

Hematochemical Analysis

The hematological analyses were performed on a Sysmex XE 2100 system (Sysmex, Japan) and the coagulation features were determined using the STAR Max analyzer (Stago Group, France); the biochemical parameters were measured on a Roche COBAS 6000 system (Roche Diagnostic, Basel, Switzerland) using Roche reagents, calibrators (Calibrator for automated systems [Cfas]/Cfas proteins), and control materials at two different levels (Precicontrol ClinChem Multi 1 and 2). All of the methods for the enzyme activity measurements were standardized to IFCC reference measurement procedures. The point of care (POC) measurements and the hemogas analysis were undertaken using Rapidpoint 500 (Siemens Healthcare).