

1 **The external validity of machine learning-based prediction scores from hematological**
2 **parameters of COVID-19: A study using hospital records from Brazil, Italy, and Western**
3 **Europe**

4

5 Ali Akbar Safdari^{1#}, Chanda Sai Keshav^{1#}, Deepanshu Mody^{1#}, Kshitij Verma^{1#}, Utsav
6 Kaushal^{1#}, Vaadeendra Kumar Burra^{1#}, Sibnath Ray², Debashree Bandyopadhyay^{1*}

7 1. Department of Biological Sciences, Birla Institute of Technology and Science, Pilani,
8 Hyderabad Campus, Hyderabad, Telangana, 500078

9 2. Gencrest Private Limited, 301-302, B-Wing, Corporate Center, Marol Pipeline Road, Andheri
10 Kurla Road, Mumbai, 400059

11 *corresponding author

12 # equal contribution (names are arranged alphabetically)

13

14

15 **Abstract:**

16 **Background**

17 The COVID-19 pandemic is the deadliest threat to humankind caused by the SARS-COV-2
18 virus in recent times. The gold standard for its detection, quantitative Real-Time Polymerase
19 Chain Reaction (qRT-PCR), has several limitations regarding experimental handling,
20 expense, and time. While the hematochemical values of routine blood tests have been
21 reported as a faster and cheaper alternative, the external validity of the model on a diverse
22 population has yet to be thoroughly investigated. Here we studied the external validity of
23 machine learning-based prediction scores from hematological parameters recorded in Brazil,
24 Italy, and Western Europe.

25 **Methods and Findings**

26 The publicly available hematological records (raw sample size (n) = 195554) from hospitals of
27 three different territories, Brazil, Italy, and Western Europe, were preprocessed to develop the
28 training, testing, and prediction cohorts for ML models. A total of eight (sub)datasets were
29 trained on seven different ML classifiers. The XGBoost classifier performed consistently better
30 on all the datasets producing eight different models. The working models include a set of either
31 four or fourteen hematological parameters. The internal performances of the XGBoost models
32 (AUC scores range from 84% to 97%) were superior to the ML models reported in the literature
33 for a few datasets (AUC scores range from 84% to 87%). The external performance (AUC
34 score) was 86% when the model was trained and tested on fourteen hematological parameters
35 obtained from the same country (Brazil) but on independent datasets. However, the external
36 performances were reduced when tested across the populations; 69% when trained on
37 datasets from Italy (n=1736) and tested on datasets from Brazil (n=602)) and 65%, when
38 trained on datasets from Italy and tested on datasets from Western Europe (n=1587))
39 respectively.

40 **Conclusion**

41 For the first time, this report showed that the models trained and tested on the same population
42 but on separate records produced reasonably accurate results. The study promises the
43 confidence of these models trained and tested within the same populations and has the
44 potential application to extend those to other demographic locations. Both four- and fourteen-
45 parameter models are publicly available; <https://covipred.bits-hyderabad.ac.in/home>

46 Author Summary:

47 COVID-19 has posed the deadliest threat to the human population in the 21st century. Timely
48 detection of the disease could save more lives. The RT-PCR test is considered the gold
49 standard for COVID-19 detection. However, there are several limitations of the technique that
50 suggests developing an alternate detection protocol that would be efficient, fast, and cheap.
51 Among several other alternate detection techniques, hematology based Machine-Learning
52 (ML) prediction is one. All the hematology-based predictions reported so far in the literature
53 were only internally validated. Considering the need to develop an alternate protocol for rapid,
54 near-accurate, and cheaper COVID-19 detection techniques, we aim to externally validate the
55 hematology-based ML prediction. Here external validation indicates use of two independent
56 datasets for model training and testing, in contrast to internal validation where the same
57 dataset splits into train and test sets. We have integrated published clinical records from Brazil,
58 Italy, and West Europe hospitals. Internal ML model performances are superior compared to
59 those reported in literature. The external model performances were equivalent to the internal
60 performances when trained and tested on the same population. However, the external
61 performances were inferior when train and test sets were from different populations. The
62 results promise the utility of these models on the same populations. However, it also warns to
63 train the model on one population and test it on another. The outcome of this work has the
64 potential for an initial screen of COVID-19 based on hematological parameters before qRT-
65 PCR tests.

66

67

68 **Introduction:**

69 The COVID-19 infection has posed the deadliest threat to the health of the human population
70 in the 21st century. Likely, the danger is far from over concerning the emerging variants of
71 COVID-19, such as alpha (B.1.1.7), beta (B.1.351), gamma (P.1), delta (B.1.617.2), lambda

72 (C.37), and omicron (B.1.1.529)¹, along with other frequently mutating respiratory diseases,
73 like, influenza virus A (H1N1)². Due to the nature of the disease, timely detection of COVID-
74 19 is of utmost importance. Hence, detection techniques play a pivotal role in its diagnosis.
75 The gold standard of COVID-19 detection is quantitative Real-Time Polymerase-Chain-
76 Reaction (qRT-PCR). This method has several limitations, like manual errors during sample
77 (nasal and oral swab) collection, operational errors, etc.³. Moreover, the time required for the
78 experiment and availability of the detection kits at a mass level becomes difficult in a vast
79 population with a large number of infections. The test is also costly for low-income groups. An
80 accurate, rapid, and low-cost prediction strategy would supplement the initial screening,
81 particularly in a country like India, with the second-largest population in the world.

82 The most common clinical feature of severe COVID-19 is pneumonia with fever, cough,
83 fatigue, headache, diarrhea, hypoxia, and dyspnoea. The latest variant, omicron, has some
84 common symptoms with the earlier SARS-COV-2 strains, although with lesser severity due to
85 mild infection in the lower respiratory tract and reduced probability of hospitalization¹. In the
86 case of mild COVID-19 infection, either no (asymptomatic) or only mild pneumonia is observed.
87 In moderate infection, dyspnoea, hypoxia, and lung injury may occur. In severe infection,
88 respiratory failure to multi-organ failure occurs. In brief, severe cases of COVID-19 can lead
89 to a systemic infection affecting almost all of the major organ systems. As a result, patients of
90 COVID-19 exhibit a wide range of hematologic abnormalities that changes with disease
91 progression, severity, and mortality⁴. For example, the white blood cells sense and respond
92 to the microbial threats⁵; blood platelet expression and platelet counts are altered^{6 7} - platelet
93 hyperactivity was demonstrated as one of the unique features of COVID-19 infection⁸. Hence,
94 a complete blood count (CBC) could serve as a biomarker for COVID-19. Screening the
95 COVID-19 infection in terms of CBC has been attempted by various research groups
96 worldwide,^{9 10 11 12 13 14}.

97 Some of these research groups used machine learning (ML) approaches to exploit the CBC
98 parameters from a specific population for disease prediction; the Area Under Curve (AUC)

99 performance ranges from 84% to 87% in those models. So far, no report is available to test
100 the applicability of the hematology-based ML models across different ethnicity and
101 populations. The combination of CBC parameters varies with ethnicity. However, some blood
102 parameters alteration, such as lymphocytopenia¹⁵ ¹⁶ leucopenia, and thrombocytopenia ¹⁷ ¹⁸
103 ¹⁹, are common due to COVID-19. In this work, we combined different hematological
104 parameters from various populations to develop the optimal ML models and tested them on
105 independent datasets obtained from other populations. Standardization across different ML
106 algorithms yields eXtreme Gradient Boost (XGBoost) as the best-performing model across the
107 datasets compared to published literature. For the first time, we report the external validity of
108 the prediction scores trained, tested, and predicted across the populations. The models
109 performed the best when trained and tested on the same population but on different records
110 (datasets).

111

112 **Method:**

113 **Description of clinical datasets for training, validation, and prediction:**

114 ***Dataset 1:***

115 Dataset-1 was generated based on anonymized patient data publicly available from Hospital
116 Israelita Albert Einstein, in São Paulo, Brazil <https://www.kaggle.com/einsteindata4u/covid19>.
117 The data were recorded from February 26th, 2020, to March 23rd, 2020. The cases and controls
118 for this dataset include the patients whose samples were collected to perform the SARS-CoV-
119 2 qRT-PCR and additional laboratory tests during a visit to the hospital.

120 The initial data set consisted of 558 positive and 5086 negative cases of COVID-19. This
121 dataset was processed to minimize the null-value columns and eliminate the negative
122 instances with many null values. The value (x_i) in each cell was pre-normalized (at the source)
123 to a mean value (μ) of zero and a unit standard deviation (σ); this was termed as 'normalized
124 count'; $x_i' = (x_i - \mu) / \sigma$. The same normalization scheme has been used throughout the

125 subsequent datasets. The columns with null values appearing more than 90% were dropped.
 126 The records (rows) showing positive results were retained by default, and the negative records
 127 were maintained only with more than 10% non-null entries. This processed dataset, termed
 128 dataset 1, contains thirty-seven features and 2004 records, 558 positives and 1446 negatives
 129 (Table 1). The negative to positive sample size ratio, 2.59, is four times less than that in the
 130 published model (11.51)⁹. Here ‘features’ refer to x-parameters used to train the model; the
 131 definition excludes the y-parameter, SARS-COV2 results (positive or negative). This definition
 132 is consistently used in the subsequent datasets. These thirty-seven features were categorized
 133 into four classes, namely, i) age, ii) severity of the infection, iii) hematological features, and iv)
 134 co-morbidities (Table S1).

135

136 Table 1: Statistics of the datasets

Dataset	No. of entries (P+N)	No. positive cases (P)	No. of negative cases (N)	Default scale_posweight (=N/P)	No. of features used
1 (i)	2004	558	1446	2.59	37
1a	602	83	519	6.25	18
1b	602	83	519	6.25	14
1c	602	83	519	6.25	4
2a (ii)	1388	765	623	0.81	31
2b	1736	816	920	1.13	4
3a (iii)	5872	1772	4100	2.31	21
3b	12105	8926	3176	0.356	14

137

I. <https://www.kaggle.com/einsteindata4u/covid19>

138

II. <https://zenodo.org/record/4081318#.X4RWqdD7TIU>

139

III. <https://repositoriodatasharingfapesp.uspdigital.usp.br/>

140

141 **Dataset 1a:** A subset of dataset 1 was curated with eighteen features – patient age quantile,
142 three hospitalization conditions, namely, patients admitted to regular ward, semi-ICU, and
143 ICU, and fourteen hematological parameters. Co-morbidities were excluded from dataset 1a.
144 The total number of records was 602, with 83 positives and 519 negatives.

145 **Dataset 1b:** A subset of dataset 1 was curated based on hematological features only. Other
146 parameters, namely, co-morbidities, patient age quantile and patient admission status, were
147 dropped in this dataset. All features with fewer than 90% of non-null values were dropped. All
148 the records that have 100% null values were dropped. The preprocessing resulted in a dataset
149 of fourteen hematological features and 602 records, 83 positives, and 519 negatives. Thus,
150 the negative-to-positive sample size ratio was 6.25.

151 **Dataset 1c:** A third subset of dataset 1 (dataset 1c) was curated from dataset 1b based on
152 four blood count features (Figure 1) that have shown a higher correlation with the qRT-PCR
153 results. The number of records, positives, and negatives are identical to dataset 1b. These
154 four blood count features were also reported as significant for SARS-COV-2 infection in
155 published literature⁹.

156

157 **Dataset 2:**

158 This dataset was obtained from San Raphael Hospital (OSR), Italy¹¹. In the original OSR
159 dataset, there were 1736 entries with a total of 72 features, and those included 36
160 hematological features. The samples were collected from patients admitted to OSR from
161 February to May 2020. Fifty-two percent of the patients were COVID-19 positive.

162

163 **Dataset 2a:** These 1736 entries were processed such that all rows (records) with more than
164 66% null values were dropped. The processed dataset contained 1388 records, 765 positives,
165 and 623 negatives. This dataset includes 31 features: age, sex, a feature for suspicion
166 (representing subjective analysis of the patient by a physician), and 28 hematological
167 parameters (Figure 1). The ratio of negative to positive records was 0.81.

168 **Dataset 2b:** A subset of dataset 2 was curated with only four blood count parameters (similar
169 to dataset 1c). No columns or rows were dropped here, as there were no rows with less than
170 66% null values. Dataset 2b has 1736 records, 816 positives, and 920 negatives.

171

172 **Dataset 3:**

173 Dataset 3 was obtained from the Covid Data Sharing initiative created by a consortium led by
174 FAPESP (Sao Paulo Research Foundation) and USP (University
175 of Sao Paulo, Brazil). The data originated from three prominent private hospitals in Sao Paulo,
176 Brazil - Fleury Institute, Sírio-Libanês Hospital, and Albert Einstein Hospital, from November 1st,
177 2019, to July 1st, 2020 (<https://repositoriodatasharingfapesp.uspdigital.usp.br/>). The data was
178 anonymized from patients tested for COVID-19 (serology or RT-PCR).

179 The raw data obtained from the data sharing initiative had multiple rows (records)
180 corresponding to individual patients containing different clinical features ("long-form" of the
181 dataset). The "long form" of the dataset was converted, using an in-house python code, to the
182 "wide form," where one row corresponds to all the clinical features of a patient. The "wide
183 form" of the dataset has 189227 records and 454 features. These 454 features were common,
184 as there were duplicates in the column headers (due to different reference ranges) for some
185 features. After deduplication, the feature number was reduced to 104.

186 **Dataset 3a:** The non-duplicated features were further filtered by excluding the following
187 conditions, i) no qRT-PCR results available, ii) all the rows with more than 66% null values,
188 and iii) the Pearson correlation of that particular feature (for the SARS-COV-2 results) less
189 than 0.05. A total of twenty-one hematological indices (features) were identified based on the
190 above cutoff (Figure 1). The final dataset contains 5872 records, 1772 positives, and 4100
191 negatives.

192 **Dataset 3b:** The deduplicated 'wide form' of the data (189227 records and 104 features) were
193 filtered with the following conditions— a) qRT-PCR results present, b) records with null values

194 less than 66%, and c) fourteen hematological parameters present, as in dataset 1b. The total
195 number of records present in the dataset was 12105 records.

196 All the processed datasets have a 90:10 split between the training and the test data.

197

198 **Description of the clinical dataset for blind prediction:**

199 ***Western European dataset:***

200 This dataset was obtained from several hospitals in Western Europe (Table S2). The dataset
201 includes the patients from the first day of hospitalization to nearly five weeks¹³. This published
202 data was in the form of twenty separate tables that we merged into a single file comprising
203 2587 entries and thirty-seven features. According to the source authors¹³, there are two
204 stages of the disease, a) early stage, from day zero through three (total of four days), and b)
205 advanced stage, comprising all the subsequent days. This blind prediction dataset includes
206 only four hematological parameters consistent with dataset-2b.

207

208 **Machine Learning (ML) approaches:**

209 The machine learning (ML) algorithms were implemented in Python (3.7.13) using the
210 following libraries, Numpy (1.21.6), Pandas (1.3.5), XGBoost (0.90), Scikit-learn (1.0.2),
211 Seaborn (0.11.2), Matplotlib (3.2.2) and Pickle 4.0 libraries.

212 *Different algorithms:*

213 The algorithm primarily employed was the Extreme Gradient Boost (XGBoost) classifier that
214 implements gradient-boosted decision trees (with enhanced speed and performance) and
215 trains a class-weighted (or cost-sensitive) version of imbalanced classification²⁰. XGBoost, a
216 ternary classifier, considers null entries one of the classes that handle the null-entry values.
217 Other classifiers tested on these datasets were logistic regression, Fischer linear discriminant
218 Naïve Bayes, SVM, random forest, and K-Nearest Neighbor (KNN). Logistic regression
219 predicts the output of a categorical dependent variable by fitting an "S" shaped logistic function
220 that indicates two maximum values, 0 or 1. Fischer linear discriminant classifier maximizes

221 the separation between the projected class means and minimizes the class overlap leading to
222 well-separated classes. Naive Bayes is a classification technique based on the Bayes theorem
223 with an assumption of independent predictors; a particular feature is independent of another
224 feature in a class. The SVM algorithm aims to create the best line or decision boundary to
225 segregate n-dimensional space into classes to accommodate a new data point. The best
226 decision boundary, a hyperplane, is made based on the extreme points (vectors). Random
227 forest is a concept of ensemble learning – a combination of multiple classifiers to solve a
228 complex problem and improve the model performance. As the name suggests, Random Forest
229 contains several decision trees on various subsets of the given dataset and takes the average
230 to improve the predictive accuracy of that dataset. KNN algorithm stores all the available data
231 and classifies a new data point based on the similarity by placing a new data point in the
232 nearest category. Thus, new data belongs to an appropriate class.

233 *Hyper-parameter used in XGBoost classifier:*

234 To normalize the imbalance in the number of negative and positive data points in the XGBoost
235 classifier, hyper-parameter – “*scale_pos_weight*”
236 <https://xgboost.readthedocs.io/en/stable/parameter.html#parameters-for-tree-boosters>, was
237 introduced. The *scale_pos_weight* value was used to scale the gradient for the positive class.
238 For example, the “*scale_pos_weight*” = 2 indicates twice the weight of the positive class
239 compared to the negative class. It also overcorrects the misclassification of the positive class.
240 The loss curve (optimized to get a better model) will be affected differently in case of positive
241 and negative entry misclassification. However, large *scale_pos_weight* can help the model
242 achieve better performance for the positive class prediction (overfitting the positive class) at
243 the cost of worse performance on the negative or both classes. Hence we have consistently
244 considered the default *scale-pos-weight* (the ratio of numbers of negative to positive entries)
245 throughout this report.

246 *Imputation for other ML models:*

247 Unlike XGBoost, most ML algorithms cannot handle null values, thus requiring data
248 imputation. We imputed missing values through the IterativeImputer module in the Scikit-learn
249 package (<https://scikit-learn.org/stable/modules/impute.html#multivariate-feature-imputation>),
250 which imputes values for null data points for each feature iteratively. It does so by fitting a
251 regressor to the other feature columns (X-parameter) for records with known values of the
252 target feature (y-parameter) and then predicts missing values of the target feature.

253 *Performance metrics:*

254 The performance metrics used were accuracy, specificity, and sensitivity, defined by true
255 positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (eq 1-3).

$$256 \text{ Accuracy} = (TP+TN)/(TP+TN+FP+FN) \dots \dots \dots \text{Eq.1}$$

$$257 \text{ Specificity} = (TN/TN+FP) \dots \dots \dots \text{Eq.2}$$

$$258 \text{ Sensitivity} = (TP/TP+FN) \dots \dots \dots \text{Eq.3}$$

259 The fourth metric was the Area Under the ROC Curve (AUC). The AUC was computed from
260 prediction scores using the roc_auc_score ([https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html)
261 [learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html)) module of the
262 sklearn—metrics library. A ROC curve (Receiver Operating Characteristic curve) plots the
263 performance (True Positive Rate (TPR) versus False Positive Rate (FPR)) of a classification
264 model at all classification thresholds. TPR is synonymous with sensitivity, also known as recall.
265 FPR is FP/(FP + TN). AUC measures the Area under ROC (as defined by TPR versus FPR)
266 curve from (0,0) to (1,1) along the x-axis (FPR axis). AUC ranges from 0 to 1; 0 implies a
267 100% wrong model, and 1 indicates a 100% correct model.

268

269 **Design of the web server:**

270 The web server hosted two different models, a four-hematological parameter model and a
271 fourteen-hematological parameter model. The web server was developed on an HTML
272 framework, with five working HTML files: a landing page and two pages each for each method,
273 one for data input and the other for prediction display. The basic skeleton of the HTML files

274 was formatted with CSS code, and these files were deployed via the python module, Flask.
275 Python libraries, like numpy and pandas, were used to collect and process the input, with the
276 responses generated by the XGBoost models.

277

278

279 **Results and Discussion:**

280 *Clinical datasets exploited for feature selection:*

281 Three independent clinical datasets (dataset 1, dataset 2, and dataset 3) were curated and
282 processed from hospitals in Brazil and Italy (Figure 2). Hematological features were selected
283 from these datasets based on the Pearson correlation coefficients computed between the
284 features and the SARS-COV-2 results (positive or negative) (Figure 3).

285 For dataset-1, four features (out of thirty-seven) showed higher correlation values (cutoff value
286 $\sim \pm 0.2$) with SARS-COV-2 results. These four features were platelet counts, monocytes,
287 eosinophils, and leukocytes (all reported in $10^9/L$). Only monocytes have shown a significant
288 increase in their values in SARS-COV-2 patients (positive correlation). The remaining
289 parameters decreased during infection (negative correlation). Careful observation revealed
290 that in the case of non-admitted patients, monocyte increase is maximum, suggesting that
291 innate immunity is handling the infection. On the other hand, platelet volume (MPV) increased,
292 and platelet counts decreased in the case of regular ward patients, clearly indicating the
293 increase in platelet size. Thus the immune system will be affected, and the number of immune
294 cells will decrease, justifying the negative correlation of eosinophil, leukocytes, and platelet
295 count with SARS-COV-2 disease. The low platelet counts accounted for severe COVID-19
296 patients, even down in non-survivors compared to the survivors²¹. The correlation coefficient
297 values between SARS-COV-2 results and different features reported elsewhere were similar
298 to this observations¹¹. Hence, dataset-1c was developed on these four features.

299 For dataset 2a, eight features (out of twenty-eight) have shown correlation values outside the
300 cutoff. Those features were i) aspartate aminotransferase, ii) lactate dehydrogenase, iii)

301 leukocyte ($10^9/L$), iv) eosinophil (%), v) basophil (%), vi) eosinophil count, vii) lymphocyte
302 count and viii) basophil count (all the counts in $10^9/L$). Two features: Aspartate
303 aminotransferase and lactate dehydrogenase have increased in COVID-19 patients. The
304 remaining other hematological features decreased. In datasets 1c and 2a, there are two
305 features, leukocyte count and eosinophil count, commonly drop with SARS-COV-2 results that
306 presumably indicate that despite variable immune response in different populations, some
307 hematological features are common in SARS-COV-2 disease across the populations.

308 For dataset 3a, four parameters, lactate dehydrogenase, partial oxygen pressure in the artery,
309 serum ferritin, and serum magnesium, have a higher positive correlation (>0.1) with SARS-
310 COV-2 results. Whereas basophil, eosinophil, leukocyte, and lymphocyte counts have a
311 higher negative correlation (<-0.1) with SARS-COV-2 results. In datasets 1c, 2a, and 3a, two
312 clinical features, leukocyte count and eosinophil count, were common.

313 *Comparative performances of seven different ML models on current datasets:*

314 Eight datasets (Figure 2) from three primary datasets, 1, 2, and 3, were derived based on
315 either higher correlation with SARS-COV-2 results or to make parity (in terms of the number
316 of features) with other datasets. The overall statistics of these eight datasets are shown (Table
317 1). Different ML models were trained on these datasets. The performances were measured
318 using the receiver operating characteristic (ROC) curves (Figure 4). XGBoost outperformed
319 other methods for all datasets except dataset 1a. The internal evaluation showed that the
320 XGBoost model outperformed all the datasets when all four performance metrics, namely,
321 accuracy, sensitivity, specificity, and AUC scores, were considered together (Table S3).

322 Datasets-1 and 1c have shown optimal performances (AUC scores 0.94 and 0.97,
323 respectively) in all four metrics. For dataset 1c, sensitivity was observed as 1.0, indicating
324 100% correct prediction of True Positive (TP) values, presumably, due to overcorrection of the
325 TP values in a small dataset ($n=602$) with a low population of positives ($n=83$), leading to large
326 *scale-pos-weight* of 6.25. As mentioned in the method section, large *scale-pos-weight*
327 improves the performance of the positive class prediction at the cost of the negative class

328 prediction. The XGBoost models, when compared with the other models, in terms of all the
329 metrics, the notable observation was low sensitivity values for dataset 1 (small sample size,
330 n=602) and allied subsets (datasets, 1a to 1c) for almost all the models except Naïve Bayes
331 classifier. The low sensitivity values for datasets 1b and 1c are presumably attributed to the
332 smaller size and shallow positive populations in those datasets. Most likely, the XGBoost,
333 being a ternary classifier, can more effectively handle the class imbalance than the imputations
334 performed in other ML methods. However, the low sensitivity problem was absent in datasets
335 2a and 2b, as the number of positives and negatives were equivalent (Table 1).

336

337 *Comparison of internal performances of the XGBoost model with published reports:*

338 The internal performances of the XGBoost model were compared with reported methods from
339 the published literature^{9 11}. The results from the XGBoost model outperformed the published
340 reports (Table 2 and Figure 5).

341

342 Table 2: Internal evaluation of the XGBoost model on different datasets and comparison with
343 published datasets

Dataset	Sensitivity	Specificity	Accuracy	AUC score	Published AUC score
1	0.826	0.974	0.940	0.941	
1a	0.875	0.925	0.918	0.967	
1b	0.750	0.887	0.869	0.922	0.87 (ref 8)
1c	1.000	0.906	0.918	0.939	0.87 (ref 8)
2a	0.830	0.843	0.835	0.906	0.84 (ref 10)
2b	0.845	0.733	0.787	0.842	
3a	0.719	0.799	0.776	0.835	
3b	0.784	0.733	0.746	0.842	

344

345 *Selection of working XGBoost models for external evaluation across the populations:*

346 As per the results, the XGBoost model performed the best on dataset 1c, having four
347 hematological parameters. However, the performance of the XGBoost model on dataset 1b,
348 having fourteen hematological parameters, was comparable to that of dataset 1c, with a
349 slightly lower AUC Score (0.94 versus 0.92). Based on these observations, we hypothesize
350 both four-parameter and fourteen-parameter models as the working ML models for COVID-19
351 testing and blind predictions across different populations. Although the internal performances
352 were the best with datasets 1a and 1c, the overfitting of the data due to small sample sizes
353 was an issue, as discussed above. Hence, we selected two other XGBoost models with four
354 and fourteen parameters obtained from datasets 2b (Italy) and 3b (Brazil), *albeit* with a slightly
355 lowered AUC score of 0.842 in both cases. These two were the final working models (training
356 dataset) for external evaluation.

357

358 *External evaluation of XGBoost models with four hematological parameters across Italian and*
359 *Brazilian populations:*

360 External evaluation for the four-parameter model was performed on the test dataset 1c from
361 Brazil. Note that the training dataset 2b was from Italy. The sensitivity was 0.81 with a lower
362 specificity value; the AUC score was 0.69 (Table 3a). For the first time, an ML model was
363 trained on one ethnic group and tested on another ethnic group with reasonably good
364 performance.

365

366 Table 3: External evaluation of XGBoost algorithm based on a) 4- hematological features and
367 b) 14-hematological features trained and tested across different datasets.

368 a)

Training set/test set	Sensitivity	Specificity	Accuracy	AUC Score

Dataset 2b (Italian) / dataset 1c (Brazilian)	0.81	0.45	0.50	0.69
--	------	------	------	------

369

370 b)

Training set/test set	Sensitivity	Specificity	Accuracy	AUC Score
Dataset 3b (Brazilian) / dataset 1b (Brazilian)	0.55	0.90	0.85	0.86

371

372 *External evaluation of XGBoost models with fourteen hematological parameters within the*
373 *Brazilian populations:*

374 The fourteen-parameter XGBoost model was trained on dataset 3b (n=12105) and tested on
375 dataset 1b (n=602), both from Brazilian populations. However, the samples in these two
376 datasets were from different time points; hence those can be considered independent data
377 sources. The AUC score for this prediction was 0.86 (Table 3b). These results were better
378 than the performance for the four-feature XGBoost model across the populations. There could
379 be multiple reasons for the better performance of the fourteen-feature model over the four-
380 feature model, a) the larger size of the training dataset, b) training and prediction data obtained
381 from the same demographic location, that is, Brazil, and c) combination of more number of
382 features with a larger dataset, presumably, yields to a better result.

383

384 *Blind prediction of XGBoost models with four hematological parameters on West European*
385 *populations:*

386 To further validate the efficacy of the working models, we have considered one more dataset
387 from published literature with thirty-seven features, including the data points along different
388 stages (time points) of COVID-19¹³. The dataset was from the literature without preprocessing
389 (no feature, records, or data points removed). According to the source authors, two distinct
390 stages of COVID-19 patients, *W.E.-early* and *W.E.-advanced*. Distributions of four
391 hematological parameters across the datasets, 1c, 2b, *W.E.-early*, and *W.E.-advanced*, were
392 compared (Figure 6). The distributions were almost the same across all the datasets for
393 Leukocytes and platelets. For eosinophils and monocytes, the distributions for datasets 2b
394 and *W.E.-early* are very similar. Moreover, distributions across datasets 1c and *W.E.-*
395 *advanced* were similar for the same features. The external performance of the model on *W.E.-*
396 *early* dataset (0.65) was high compared to that on *W.E.-advanced* dataset (0.52) (Table 4).
397 To note, *W.E.-early* and *W.E.-advanced* datasets contain information only from COVID-19
398 patients and no negative controls. Hence, only the sensitivity metric was reported (Table 4).

399

400 Table 4: Blind prediction of XGBoost model trained on dataset 2b and tested on *W.E.-early*
401 and *W.E.-advanced* datasets. The *early* and *advanced* datasets contain only COVID-19-
402 positive patient results; no negatives were available. Hence, only sensitivity values reported

Training set/test set	Sensitivity
Dataset 2b/ <i>W.E.-early</i>	0.65
Dataset 2b/ <i>W.E.-advanced</i>	0.52

403

404 *Deployment of Prediction server:*

405 We deployed a web server where two sets of inputs are acceptable for binary COVID-19
406 prediction, i) four hematological parameters (leukocyte, monocyte, eosinophil, and platelet
407 count) and ii) fourteen-parameter models in the following URL link, <https://covipred.bits->

408 hyderabad.ac.in/home. Different pages on the webserver are shown (Figure 7). The server
409 outputs the COVID-9 results, either positive or negative, with the COVID-19 probability
410 reported in percentage.

411 *Conclusion:*

412 Considering the need to develop an alternate protocol for rapid, near-accurate, and cheaper
413 COVID-19 detection techniques, we aimed to externally validate the hematology-based ML
414 prediction reported in the literature with internal evaluation only. We have integrated published
415 clinical records from Brazil, Italy, and West Europe hospitals. The data from Brazil and Italy
416 were classified into eight datasets and trained on seven different ML methods; the XGBoost
417 algorithm was the best. The internal performances of the XGBoost models were better than
418 the published reports on the same datasets. Four and fourteen-parameter XGBoost models
419 were selected for external evaluations. The external performance of the fourteen-parameter
420 XGBoost model trained and tested on the Brazilian dataset was similar to that of the internal
421 performance. However, the external performances of the four-parameter XGBoost model
422 trained on the Italian dataset and tested on a) Brazilian and b) West European datasets were
423 poorer than the previous one. The results promise the utility of these models when trained and
424 tested on the same populations. However, it also warns to use the model, with caution, trained
425 on one population and test on another. The outcome of this work has the potential for an initial
426 screen of COVID-19 based on hematological parameters before qRT-PCR tests. In future
427 work, we aim to train and test those on the Indian population to use at the healthcare centers
428 of India.

429

430

431 ***Funding Information: DB gratefully acknowledges DST-MATRICS (COVID-19 special***
432 ***call) Govt. of India, Grant/Award number: MSC/2020/000498, for funding this project.***

433 **AAS acknowledges CSIR, India, for Junior and Senior Research Fellowships, Award**
434 **Number: 09/1026(0033)/2020-EMR-I**

435

436 References:

- 437 (1) Menni, C.; Valdes, A. M.; Polidori, L.; Antonelli, M.; Penamakuri, S.; Nogal, A.; Louca,
438 P.; May, A.; Figueiredo, J. C.; Hu, C.; Molteni, E.; Canas, L.; Österdahl, M. F.; Modat,
439 M.; Sudre, C. H.; Fox, B.; Hammers, A.; Wolf, J.; Capdevila, J.; Chan, A. T.; David, S.
440 P.; Steves, C. J.; Ourselin, S.; Spector, T. D. Symptom Prevalence, Duration, and
441 Risk of Hospital Admission in Individuals Infected with SARS-CoV-2 during Periods of
442 Omicron and Delta Variant Dominance: A Prospective Observational Study from the
443 ZOE COVID Study. *www.thelancet.com* **2022**, 399 (10335), 1618–1624.
444 [https://doi.org/10.1016/S0140-6736\(22\)00327-0](https://doi.org/10.1016/S0140-6736(22)00327-0).
- 445 (2) Lubna, S.; Chinta, S.; Burra, P.; Vedantham, K.; Ray, S.; Bandyopadhyay, D. New
446 Substitutions on NS1 Protein from Influenza A (H1N1) Virus: Bioinformatics Analyses
447 of Indian Strains Isolated from 2009 to 2020. *Heal. Sci. Reports* **2022**, 5 (3), e626.
448 <https://doi.org/10.1002/hsr2.626>.
- 449 (3) Syal, K. Guidelines on Newly Identified Limitations of Diagnostic Tools for COVID-19
450 and Consequences. *J. Med. Virol.* **2021**, 93 (4), 1837–1842.
451 <https://doi.org/10.1002/JMV.26673>.
- 452 (4) Taj, S.; Kashif, A.; Arzinda Fatima, S.; Imran, S.; Lone, A.; Ahmed, Q. Role of
453 Hematological Parameters in the Stratification of COVID-19 Disease Severity. *Ann.*
454 *Med. Surg.* **2021**, 62, 68–72. <https://doi.org/10.1016/J.AMSU.2020.12.035>.
- 455 (5) Beadling, C.; Silfka, M. M. No TQuantifying Viable Virus-Specific T Cells without a
456 Priori Knowledge of Fine Epitope Specificitytitle. *Nat. Med.* **2006**, 12, 1208–1212.
457 <https://doi.org/https://doi.org/10.1038/nm1413>.
- 458 (6) Kanth Manne, B.; Denorme, F.; Middleton, E. A.; Portier, I.; Rowley, J. W.; Stubben,
459 C.; Petrey, A. C.; Tolley, N. D.; Guo, L.; Cody, M.; Weyrich, A. S.; Yost, C. C.;

- 460 Rondina, M. T.; Campbell, R. A. Platelet Gene Expression and Function in Patients
461 with COVID-19. *Blood* **2020**, 136 (11), 1317–1329.
462 <https://doi.org/10.1182/blood.2020007214>.
- 463 (7) Güçlü, E.; Kocayiğit, H.; Okan, H. D.; Erkorkmaz, U.; Yürümez, Y.; Yaylacı, S.;
464 Koroglu, M.; Uzun, C.; Karabay, O. Effect of COVID-19 on Platelet Count and Its
465 Indices. *Rev. Assoc. Med. Bras.* **2020**, 66 (8), 1122–1127.
466 <https://doi.org/10.1590/1806-9282.66.8.1122>.
- 467 (8) Comar, S. P. et. al. *COVID-19 Induces A Hyperactive Phenotype in Circulating*
468 *Platelets*; 2020. [https://doi.org/doi: https://doi.org/10.1101/2020.07.24.20156240](https://doi.org/10.1101/2020.07.24.20156240).
- 469 (9) Banerjee, A.; Ray, S.; Vorselaars, B.; Kitson, J.; Mamalakis, M.; Weeks, S.; Baker, M.;
470 Mackenzie, L. S. Use of Machine Learning and Artificial Intelligence to Predict SARS-
471 CoV-2 Infection from Full Blood Counts in a Population. *Int. Immunopharmacol.* **2020**,
472 86 (May), 106705. <https://doi.org/10.1016/j.intimp.2020.106705>.
- 473 (10) Djakpo, D. K.; Wang, Z.; Zhang, R.; Chen, X.; Chen, P.; Ketisha Antoine, M. M. L.
474 Blood Routine Test in Mild and Common 2019 Coronavirus (COVID-19) Patients.
475 *Biosci. Rep.* **2020**, 40 (8), 1–5. <https://doi.org/10.1042/BSR20200817>.
- 476 (11) Federico, C.; Andrea, C.; Davide, F.; Chiara, D. R.; Daniele, C.; Eleonora, S.;
477 Alessandra, C.; Elena, D. V.; Giuseppe, B.; Massimo, L.; Anna, C. Development,
478 Evaluation, and Validation of Machine Learning Models for COVID-19 Detection
479 Based on Routine Blood Tests. *Clin. Chem. Lab. Med.* **2020**, 59 (2), 421–431.
480 <https://doi.org/10.1101/2020.10.02.20205070>.
- 481 (12) Brinati, D.; Campagner, A.; Ferrari, D.; Locatelli, M.; Banfi, G.; Cabitza, F. Detection of
482 COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility
483 Study. *J. Med. Syst.* **2020**, 44, 135. <https://doi.org/10.1101/2020.04.22.20075143>.
- 484 (13) Linssen, J.; Ermens, A.; Berrevoets, M.; Seghezzi, M.; Previtali, G.; van der Sar-Van
485 der Brugge, S.; Russcher, H.; Verbon, A.; Gillis, J.; Riedl, J.; de Jongh, E.; Saker, J.;
486 Münster, M.; Munnix, I. C. A.; Dofferhoff, A.; Scharnhorst, V.; Ammerlaan, H.;

- 487 Deiteren, K.; Bakker, S. J. L.; van Pelt, L. J.; de Hingh, Y. K.; Leers, M. P. G.; van der
488 Ven, A. A Novel Haemocytometric COVID-19 Prognostic Score Developed and
489 Validated in an Observational Multicentre European Hospital-Based Study. *Elife* **2020**,
490 *9* (e63195), 1–28. <https://doi.org/10.1101/2020.09.27.20202168>.
- 491 (14) Abdullah, I.; Cornelissen, H. M.; Musekwa, E.; Zemlin, A.; Jalavu, T.; Mashigo, N.;
492 Chetty, C.; Nkosi, N.; Chapanduka, Z. C. Hematological Findings in Adult Patients
493 with SARS CoV- 2 Infection at Tygerberg Hospital Cape Town South Africa. *Heal.*
494 *Sci. Reports* **2022**, *5* (3), 1–9. <https://doi.org/10.1002/hsr2.550>.
- 495 (15) Guan, W.; Ni, Z.; Hu, Y.; Liang, W.; Ou, C.; He, J.; Liu, L.; Shan, H.; Lei, C.; Hui, D. S.
496 C.; Du, B.; Li, L.; Zeng, G.; Yuen, K.-Y.; Chen, R.; Tang, C.; Wang, T.; Chen, P.;
497 Xiang, J.; Li, S.; Wang, J.; Liang, Z.; Peng, Y.; Wei, L.; Liu, Y.; Hu, Y.; Peng, P.;
498 Wang, J.; Liu, J.; Chen, Z.; Li, G.; Zheng, Z.; Qiu, S.; Luo, J.; Ye, C.; Zhu, S.; Zhong,
499 N. Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.*
500 **2020**, *382* (18), 1708–1720. <https://doi.org/10.1056/nejmoa2002032>.
- 501 (16) Chong, V. C. L.; Lim, K. G. E.; Fan, B. E.; Chan, S. S. W.; Ong, K. H.; Kuperan, P.
502 Reactive Lymphocytes in Patients with COVID- 19. *Br. J. Haematol.* **2020**, *189* (5),
503 844–844. <https://doi.org/10.1111/bjh.16690>.
- 504 (17) Fan, B. E.; Chong, V. C. L.; Chan, S. S. W.; Lim, G. H.; Lim, K. G. E.; Tan, G. B.;
505 Mucheli, S. S.; Kuperan, P.; Ong, K. H. Hematologic Parameters in Patients with
506 COVID-19 Infection. *Am. J. Hematol.* **2020**, *95* (6), E131–E134.
507 <https://doi.org/10.1002/ajh.25774>.
- 508 (18) Henry, B. M.; De Oliveira, M. H. S.; Benoit, S.; Plebani, M.; Lippi, G. Hematologic,
509 Biochemical and Immune Biomarker Abnormalities Associated with Severe Illness
510 and Mortality in Coronavirus Disease 2019 (COVID-19): A Meta-Analysis. *Clinical*
511 *Chemistry and Laboratory Medicine*. De Gruyter June 2020, pp 1021–1028.
512 <https://doi.org/10.1515/cclm-2020-0369>.
- 513 (19) Jiang, S.; Huang, Q.; Xie, W.; Lv, C.; Quan, X. The Association between Severe

514 COVID-19 and Low Platelet Count: Evidence from 31 Observational Studies Involving
515 7613 Participants. *Br. J. Haematol.* **2020**, *190* (1), e29–e33.

516 <https://doi.org/10.1111/bjh.16817>.

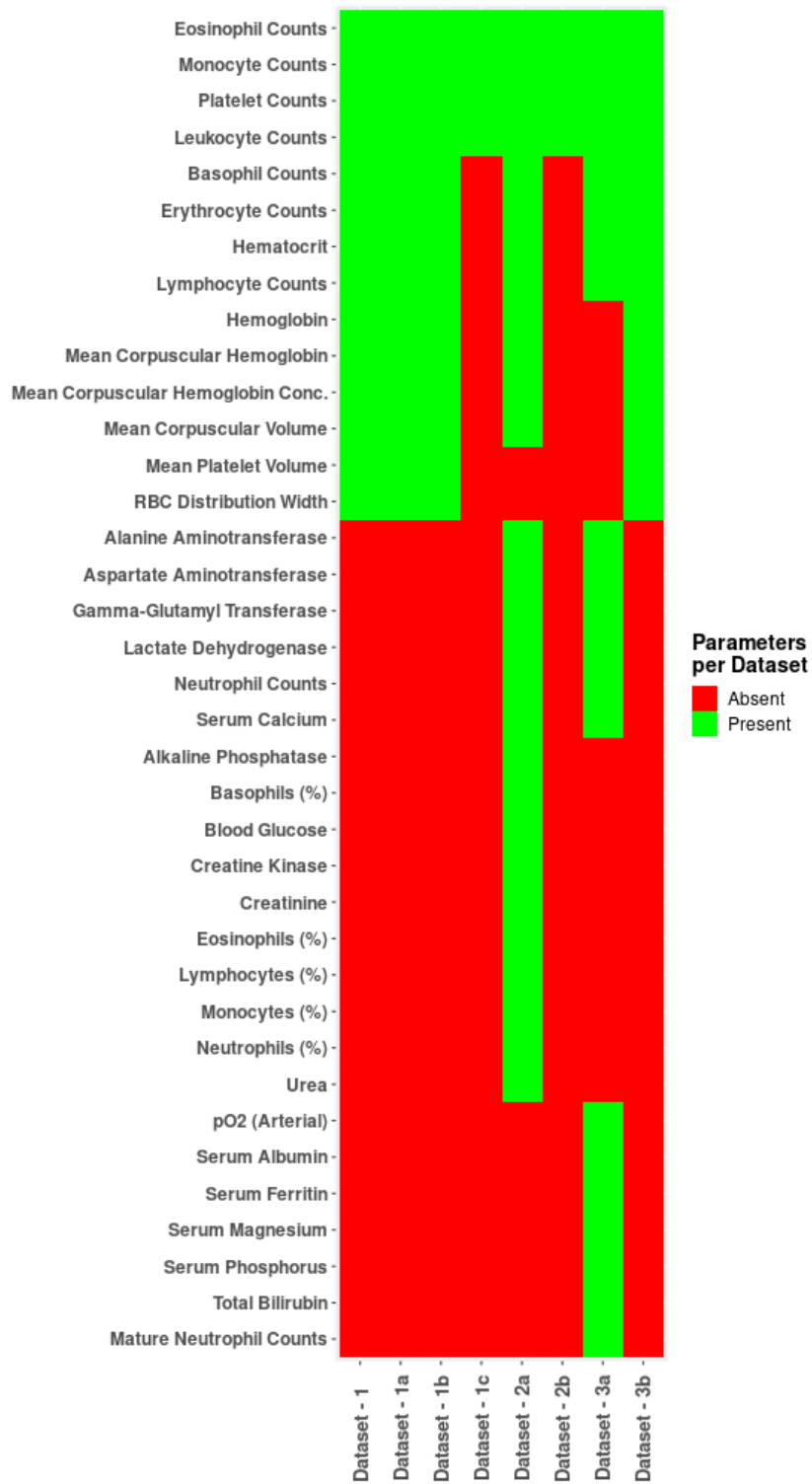
517 (20) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. ACM*
518 *SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2016**, *13-17-Aug*, 785–794.

519 <https://doi.org/10.1145/2939672.2939785>.

520 (21) Wool, G. D.; Miller, J. L. The Impact of COVID-19 Disease on Platelets and
521 Coagulation. *Pathobiology* **2021**, *88* (1), 15–27. <https://doi.org/10.1159/000512007>.

522 Figure 1: Haematological features used in different datasets. The green colour indicated the
523 presence of a particular feature in a dataset and the red colour indicated its absence.

524



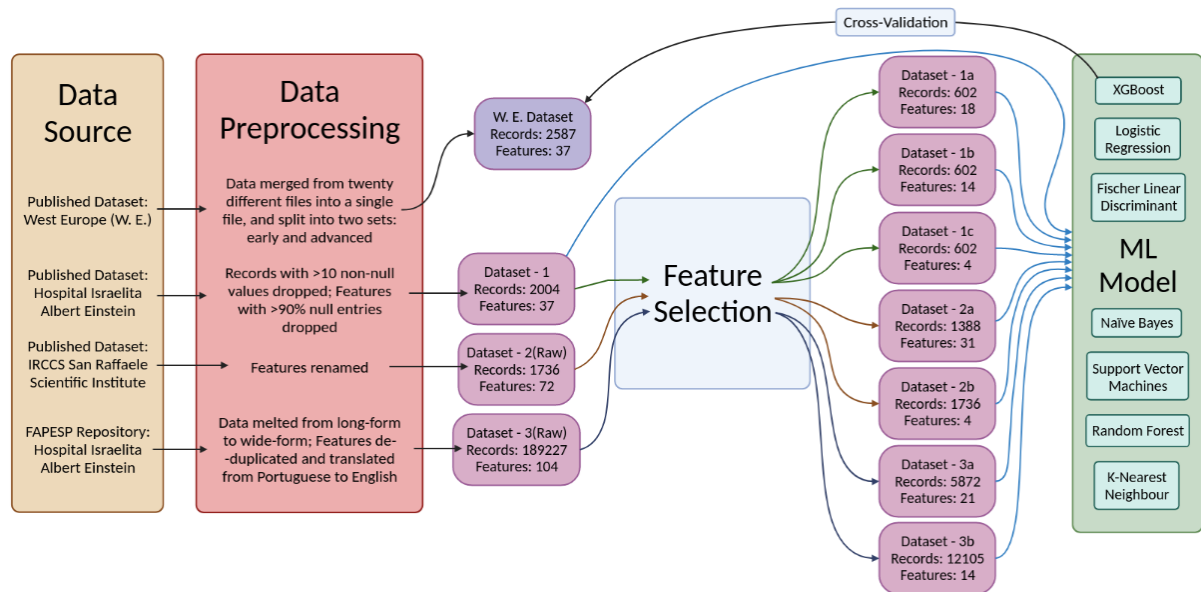
525

526

527 Figure 2: Description of data sources used for training and prediction of different ML-models

528 based on haematological features for COVID-19 characterization

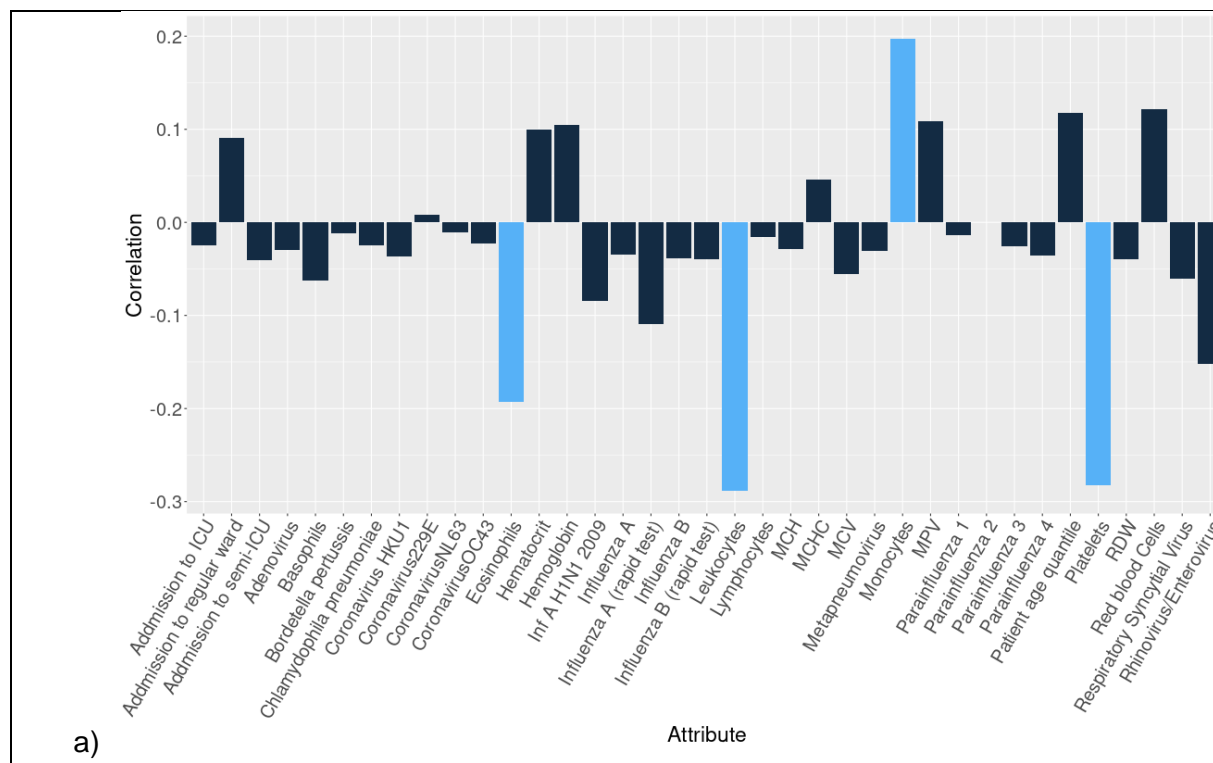
529

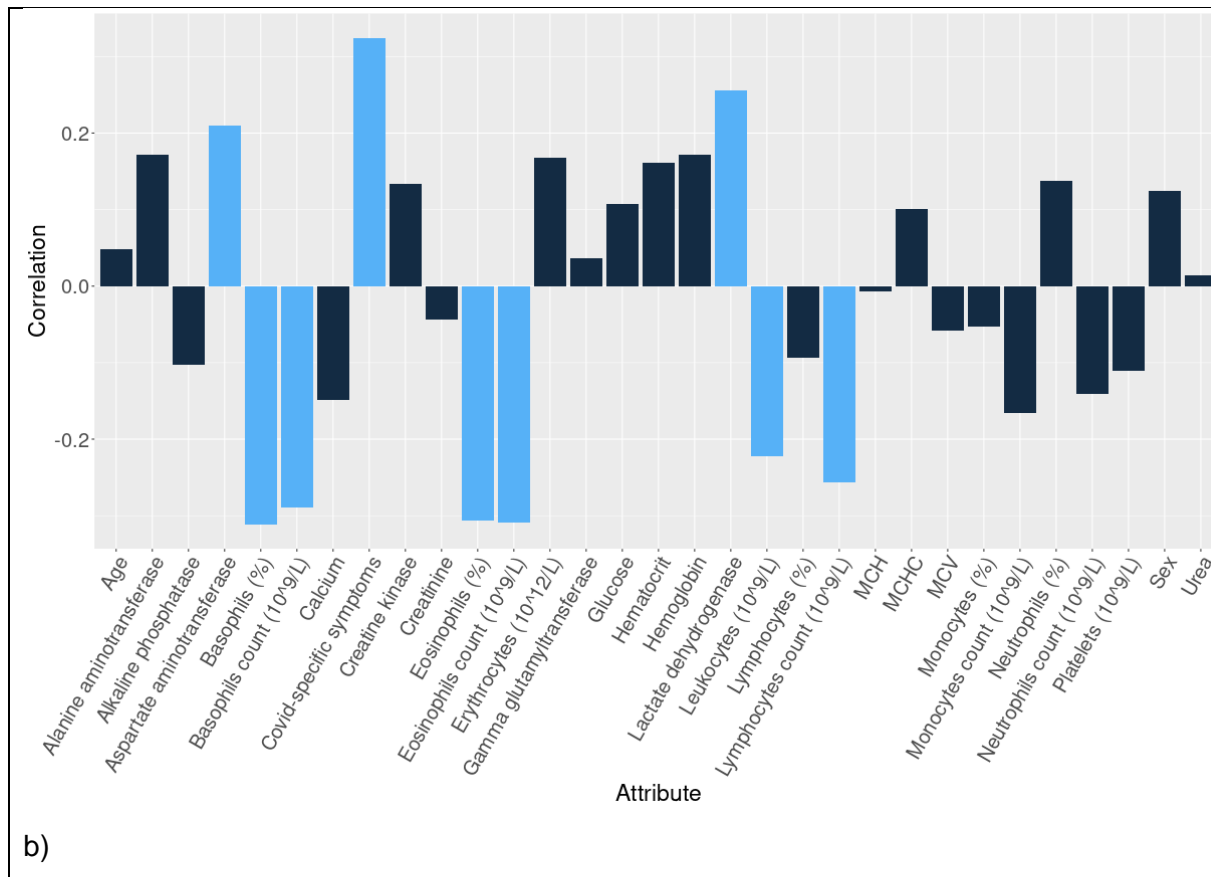


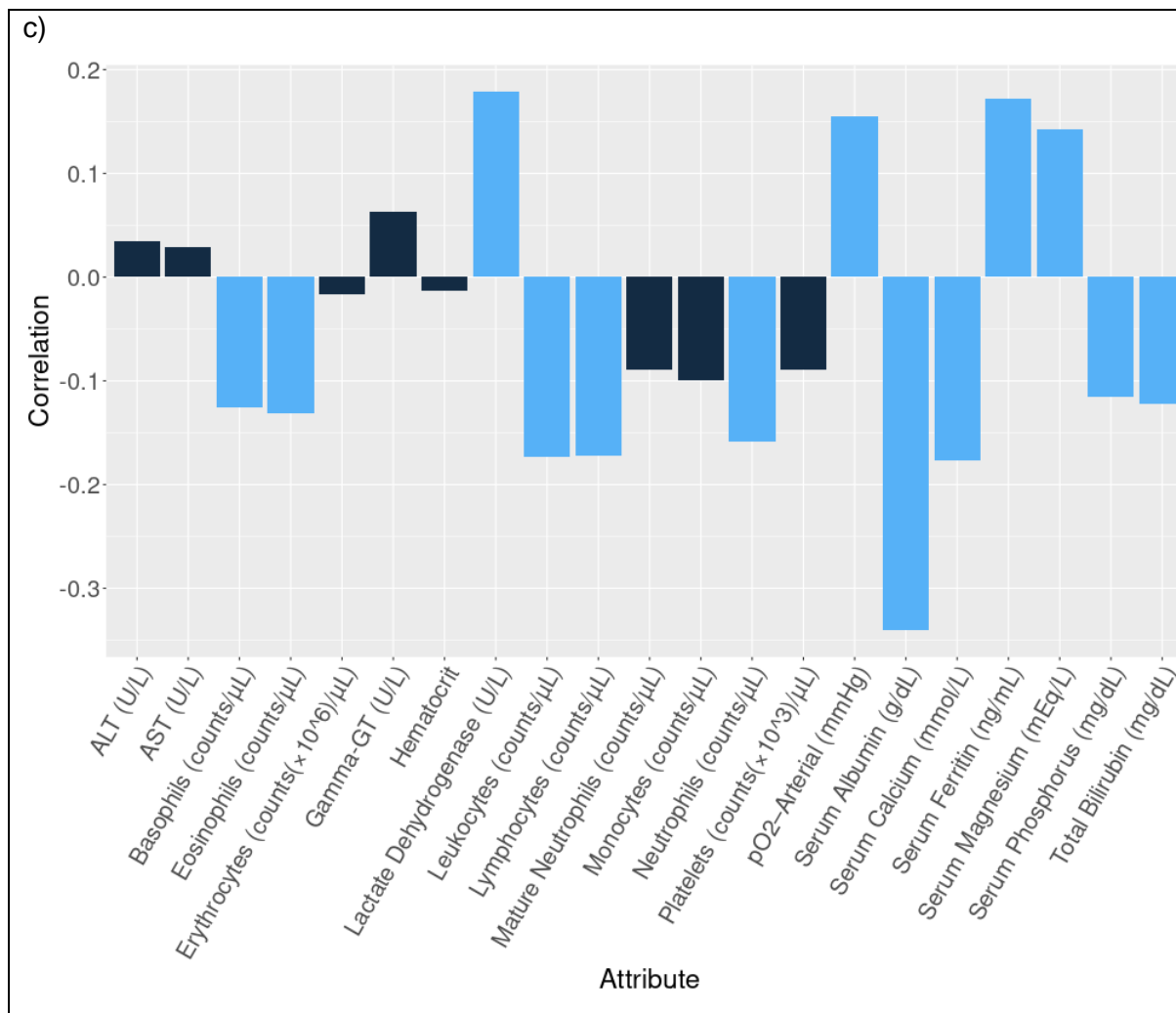
530

531

532 Figure 3: Pearson correlation coefficients between SARS-COV-2 results and individual
 533 features for a) dataset 1 b) dataset 2a and c) dataset 3a. Parameters with higher
 534 correlation ($> \sim \pm 0.2$) are shown in blue, remaining values in black, with an exception for
 535 dataset 3a (correlation cut off ± 0.1).





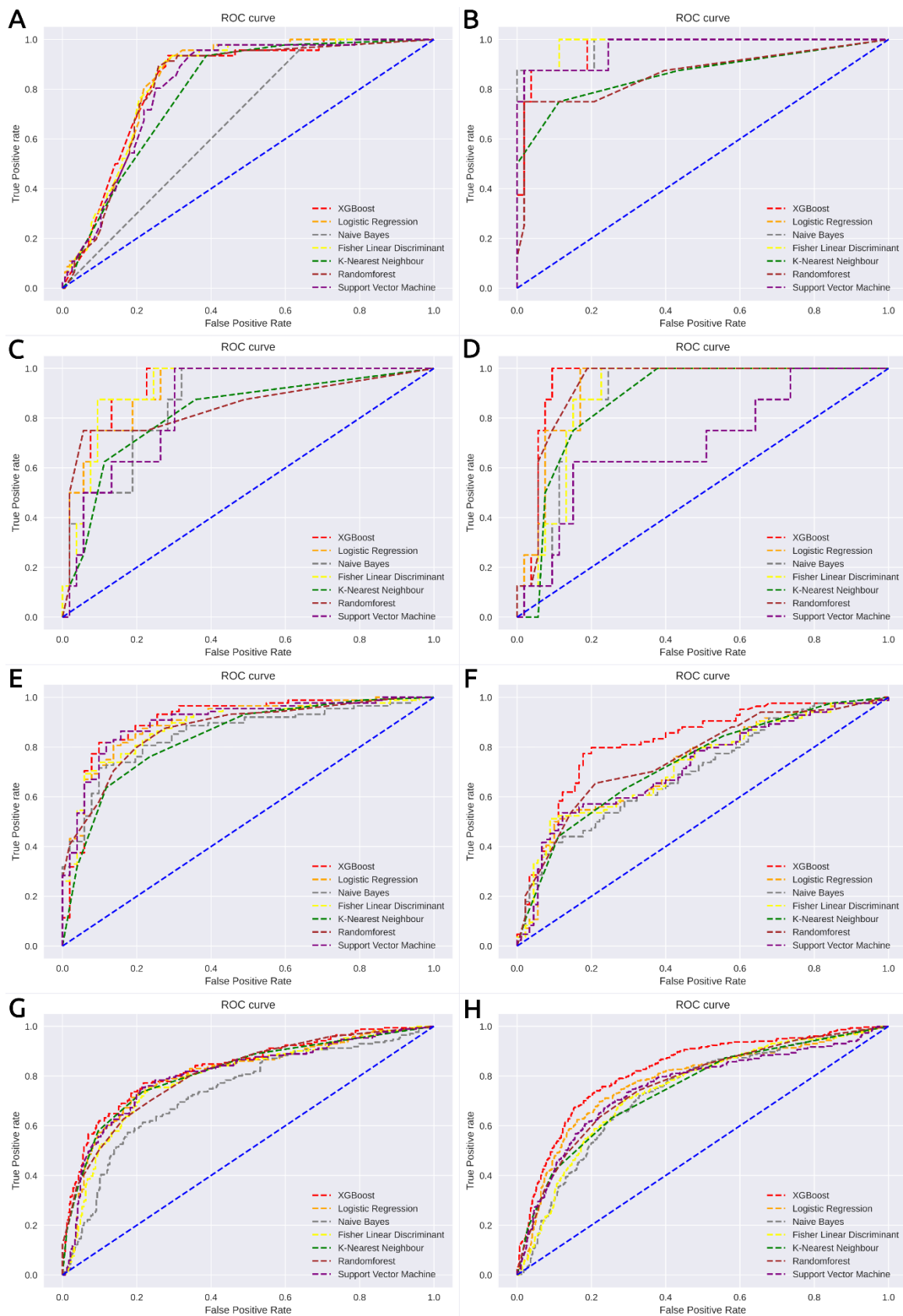


536

537

538 Figure 4: Receiver Operating Characteristics Curves (ROC) across different ML models for a)
539 Dataset 1 b) Dataset 1a, c) Dataset 1b, d) Dataset 1c, e) Dataset 2a, f) Dataset 2b, g) Dataset
540 3a and h) Dataset 3b

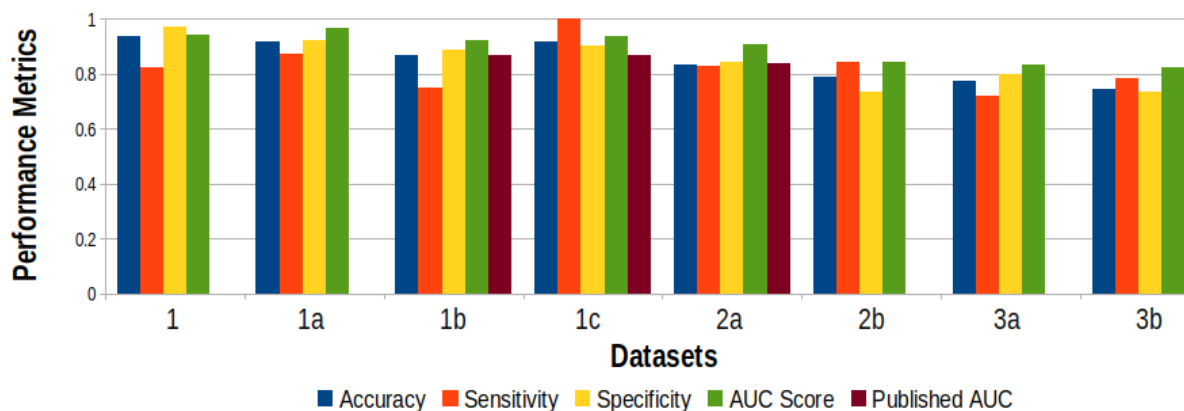
541



542

543

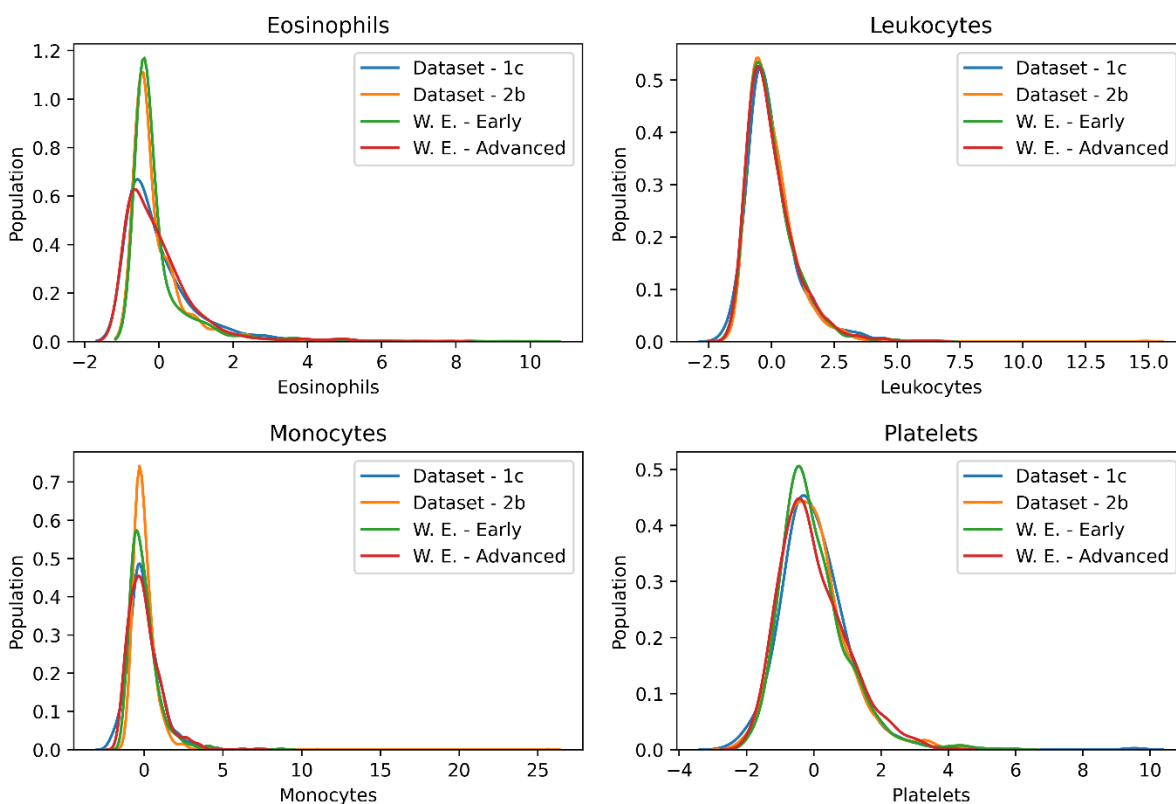
544 Figure 5: Comparative performances of different datasets trained on XGBoost model. The
 545 datasets with published AUC scores are shown in brown bars for the following datasets (1b
 546 and 1c)⁹ and 2a¹¹.



547

548

549 Figure 6: Distributions of four hematological parameters across four different datasets (two
 550 training datasets – Dataset 1c and Dataset 2b and two test datasets –*early* and *advance*). The
 551 hematological parameters are – a) platelet, b) leukocyte c) eosinophil and d) monocyte. These
 552 distributions indicate the proximity of the individual test datasets to the training datasets



553

554 Figure 7: COVID-19 prediction server based on hematological parameters, a) home page b)
 555 4-parameter prediction model and c) 14-parameter prediction model

556

a)

b)

c)

557

