Supplementary Materials

Predicting Severe COVID-19 Requiring Hospitalization with Machine Learning: A Comparison of Four Classifiers and SHAP Explainability on a Large Population Dataset

Suhaan Thayyil
Marvin Ridge High School, Waxhaw, NC

Supplementary Table S1. Data dictionary

Field-level description and timing of each predictor relative to the prediction horizon (date of confirmed positive COVID-19 test). Binary fields use the original Mexican Ministry of Health coding (1 = condition present, 2 = condition absent) except SEX where 1 = female and 2 = male. Values 97, 98, and 99 indicate missing data and were imputed as described in Methods.

Feature	Type	Available at horizon	Notes
AGE	Continuous	Yes	Patient age in years; recorded at intake.
SEX	Binary	Yes	1 = female, 2 = male; recorded at intake.
PNEUMONIA	Binary	Partial	1 = present, 2 = absent. Diagnosis sometimes confirmed at or after admission. Sensitivity analysis without this feature is reported.
DIABETES	Binary	Yes	Pre-existing diagnosis recorded at intake.
ASTHMA	Binary	Yes	Pre-existing diagnosis recorded at intake.
HIPERTENSION	Binary	Yes	Pre-existing hypertension diagnosis recorded at intake.
OBESITY	Binary	Yes	Pre-existing diagnosis recorded at intake.
CARDIOVASCULAR	Binary	Yes	Pre-existing cardiovascular disease recorded at intake.
RENAL_CHRONIC	Binary	Yes	Pre-existing chronic renal disease recorded at intake.
TOBACCO	Binary	Yes	Smoker status recorded at intake.

Supplementary Table S2. Selected hyperparameters

Hyperparameters selected by 5-fold stratified grid search cross-validation on the 80% training partition (20,000 patients), optimizing ROC-AUC. The XGBoost grid was searched with a 40-iteration randomized search to keep run time tractable. CV ROC-AUC is the mean ROC-AUC across the five folds. All models used random_state = 42.

Model	Selected hyperparameters	CV ROC-AUC
Logistic Regression	penalty = L1, C = 0.1, solver = liblinear, class_weight = balanced, max_iter = 1000	0.8938
Random Forest	n_estimators = 500, max_depth = 10, min_samples_split = 10, max_features = sqrt, bootstrap = True, class_weight = balanced	0.8929

XGBoost	n_estimators = 100, learning_rate = 0.05, max_depth = 5, subsample = 1.0, colsample_bytree = 1.0, gamma = 0, reg_lambda = 1, reg_alpha = 0	0.8933
SVM (RBF)	C = 0.1, gamma = 0.01, probability = True	0.8880

Supplementary Table S3. Subsample distribution check

Comparison of marginal distributions between the full filtered set (n = 391,979) and the 25,000-patient stratified subsample. Kolmogorov-Smirnov tests were used for all features. All p-values exceed 0.18 (most exceed 0.99), indicating no detectable distribution shift introduced by subsampling. The largest absolute mean difference across the ten features is 0.12 years for age.

Feature	Full mean	Subsample mean	Abs diff	KS stat	KS p-value
AGE	45.1703	45.0455	0.1248	0.0071	0.1897
SEX	1.5344	1.5344	0.0000	0.0000	1.0000
PNEUMONIA	1.7805	1.7829	0.0024	0.0024	0.9991
DIABETES	1.8410	1.8439	0.0029	0.0029	0.9880
ASTHMA	1.9733	1.9727	0.0006	0.0006	1.0000
HIPERTENSION	1.8036	1.8074	0.0039	0.0039	0.8769
OBESITY	1.8137	1.8164	0.0027	0.0027	0.9951
CARDIOVASCULAR	1.9782	1.9782	0.0001	0.0001	1.0000
RENAL_CHRONIC	1.9796	1.9804	0.0007	0.0007	1.0000
TOBACCO	1.9264	1.9270	0.0006	0.0006	1.0000

Supplementary Table S4. Sensitivity analysis without pneumonia

All four models were retrained with pneumonia removed from the feature set, using the same selected hyperparameters from the full-feature run. Performance dropped substantially across all four models. ROC-AUC fell to approximately 0.78, and F1-score fell from approximately 0.76 to approximately 0.59. This sensitivity analysis addresses the temporal-leakage concern: when restricted to predictors unambiguously available at the prediction horizon, the model still discriminates above chance but with markedly reduced accuracy.

Model (no pneumonia)	Test Acc.	Precision	Recall	F1	ROC-AUC	Brier
Logistic Regression	73.0%	0.518	0.701	0.596	0.784	0.189
Random Forest	73.3%	0.524	0.659	0.584	0.778	0.184
XGBoost	73.2%	0.522	0.679	0.590	0.783	0.183
SVM	72.5%	0.512	0.706	0.593	0.782	0.189

Supplementary Table S5. Pairwise DeLong tests

Pairwise DeLong tests on the held-out test set (n = 5,000). Three of the four main classifiers (Logistic Regression, Random Forest, XGBoost) are statistically indistinguishable from each other in terms of ROC-AUC. SVM trails the other three significantly. The minimal logistic regression (using only age and pneumonia) is statistically below the three top models but only by

0.005-0.007 AUC; the difference is significant in absolute terms but small in effect size, and the minimal LR is statistically indistinguishable from SVM.

Comparison	Z statistic	p-value	Significant (alpha = 0.05)
LR vs Random Forest	-0.66	0.51	No
LR vs XGBoost	-0.83	0.41	No
LR vs SVM	+3.06	0.002	Yes (LR > SVM)
Random Forest vs XGBoost	-0.28	0.78	No
Random Forest vs SVM	+3.33	<0.001	Yes (RF > SVM)
XGBoost vs SVM	+3.43	<0.001	Yes (XGB > SVM)
LR vs Minimal LR	+3.77	<0.001	Yes (full LR > minimal LR)
Random Forest vs Minimal LR	+3.21	0.001	Yes (RF > minimal LR)
XGBoost vs Minimal LR	+3.60	<0.001	Yes (XGB > minimal LR)
SVM vs Minimal LR	-0.37	0.71	No

Supplementary Note S1. Data and code availability

All analysis code, including the full revision rerun script, the 25,000-row stratified subsample patient-ID list, and the configuration used for hyperparameter tuning, is available at https://github.com/suhaanthayyil/Covid-19. The Mexico COVID-19 open dataset is publicly available from the Kaggle distribution at https://www.kaggle.com/datasets/meirnizri/covid19-dataset. Random seeds were fixed at 42 across NumPy, scikit-learn, and XGBoost. Re-running rerun_ml.py against the public dataset reproduces every number reported in the manuscript and supplementary tables.