Prediction of severe COVID-19 infection at the time of testing: A machine learning approach

Faraz Khoshbakhtian[1], Ardian Lagman[1], Dionne M. Aleman[1], Randy Giffen[2], and Proton Rahman[3]

[1]Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, Ontario, Canada
[2]IBM Canada, Toronto, ON, Canada
[3]Eastern Health, Newfoundland & Labrador, Canada

October 14, 2021

Abstract

Early and effective detection of severe infection cases during a pandemic can significantly help patient prognosis and resource allocation. We develop a machine learning framework for detecting severe COVID-19 cases at the time of RT-PCR testing. We retrospectively studied 988 patients from a small Canadian province that tested positive for SARS-CoV-2 where 42 (4%) cases were *at-risk* (i.e., resulted in hospitalization, admission to ICU, or death), and 8 ($< 1\%$) cases resulted in death. The limited information available at the time of RT-PCR testing included age, comorbidities, and patients' reported symptoms, totaling 27 features. Vaccination status was unavailable. Due to the severe class imbalance and small dataset size, we formulated the problem of detecting severe COVID as anomaly detection and applied three models: one-class support vector machine (OCSVM), weight-adjusted XGBoost, and weight-adjusted AdaBoost. The OCSVM was the best performing model for detecting the deceased cases with an average 95% true positive rate (TPR) and 27.2% false positive rate (FPR). Meanwhile, the XGBoost provided the best performance for detecting the at-risk cases with an average 96.2% TPR and 19% FPR. In addition, we developed a novel extension to SHAP interpretability to explain the outputs from the models. In agreement with conventional knowledge, we found that comorbidities were influential in predicting severity, however, we also found that symptoms were generally more influential, noting that machine learning combines all available data and is not a single-variate statistical analysis.

Contents

1 Introduction

The COVID pandemic showed that even the most resilient national healthcare systems are prone to quickly becoming overwhelmed in times of widespread infectious disease outbreaks, especially when there is no known effective treatment for the disease [9, 16, 22]. Early detection of severe COVID cases, that is, those resulting in hospitalization, ICU, or death, can help plan and manage scarce resources to best treat patients and manage scarce health care resources. We developed machine learning tools that can detect severe COVID cases at the time of RT-PCR testing, before patients present at emergency departments, are seen by any clinical practitioners, or have any diagnostic analysis performed. Outcome prediction at this earliest possible moment of knowledge of COVID infection is unique among the literature predicting COVID severity, which all make predictions after a patient has already presented at an emergency department. Additionally, our focus on a small regional area may help avoid underlying biases in the dataset and ensure the predictive models are tuned for the actual patient population being treated.

Many studies have shown the promise of machine learning in augmenting decision-making capabilities for clinicians in diagnosing COVID through imaging data [17, 25, 30] and tabular data [1, 29]. The use cases range from detecting COVID positive cases to detecting severe

cases from the infected population. To the best of our knowledge, machine learning applications for COVID prognosis have exclusively focused on supervised classification methods. Even though the reported metrics show high efficacy, they generally require somewhat large labeled datasets—18000 patients in Subudhi et al. [29] and 6995 in Assaf et al. [1]—and all rely on blood test and/or imaging data not available until a patient has already presented to an emergency department or an out-patient clinic. In the case of a global pandemic, large labeled datasets indicate human loss and missed opportunities to save lives, as well as are usually not available for single centers or individual regions to analyze and make their own predictive models. Further, datasets aggregated over numerous hospitals run the risk of biasing predictions against actual patient populations seen in smaller regional areas.

Model interpretability and decision explainability are essential parts of any health-oriented machine learning framework that aims for clinical relevance. As machine learning models become more complex, it becomes harder to explain and fully understand the reasons behind their decisions. However, especially in medical applications, it is crucial not only to verify the performance of the models but also to make sure the decisions made by them are explainable and derived from relevant pieces of information. On the one hand, biases can quickly propagate from datasets to models and onto the medical institutions. On the other hand, mere outcome predictions cannot help medical practitioners save their patients' lives. Feature interpretability and output explainability allow for assessing models' understanding of risk factors, which gives medical practitioners insight to alleviate those risk factors.

Thus, we develop an unsupervised anomaly detection and imbalance-aware supervised avenues for early risk assessment and prognosis of COVID patients, and use interpretability techniques to explain the impact of patient features on the severity predictions. Given the relative rarity of severe COVID outcomes among all COVID-positive patients, we formulate the problem of detecting severe COVID as one of anomaly detection. We train two supervised models—weight-adjusted XGBoost and AdaBoost—and one unsupervised model—a one-class support vector machine (OCSVM)—to detect at-risk cases (hospitalization, ICU, death) as well as deceased cases. Despite the limited data available at the time of RT-PCR testing, which includes age, comorbidities, patient-reported symptoms (but not vaccination status), our models show high sensitivity to the at-risk patients even when trained on a relatively small dataset. We use SHapley Additive exPlanations (SHAP) [21] to interpret model decisions based on understandable and intuitive features and feature groups, allowing for validation by clinicians. We novelly extend SHAP calculations to the particular needs of our dataset and modeling approaches, which include feature groups in addition to individual features (e.g., a "symptoms group" feature in addition to individual specific symptoms) and aggregate SHAP analysis across multiple cross-fold validations, as needed by our small dataset. Our framework can be used in the early and later stages of regional outbreaks to closely monitor the at-risk patients and plan for resource allocation.

2 Literature review

Since the start of the COVID pandemic in early 2020, a myriad of works have shown that machine learning can be an essential and effective tool in alleviating the risks and costs of a global pandemic. Both structured and unstructured data can be used to diagnose patients and/or perform prognostic assessments at different stages of their infection. These frameworks can help inform decision-making for individual COVID patients at the micro-level or integrate into the macro-level planning and early alert infrastructure.

Some studies take advantage of unstructured data such as radiographic images. These studies mostly use either chest X-rays (CXR) [17, 25, 30] or CT scans [20]; however, CXR images remain a more viable option because of their low cost and widespread adoption. Wang et al. [30] proposed COVID-net, a custom convolutional neural network (CNN) architecture for the classification of COVID infections based on CXR radiography images and report 98.9% positive predictive value (PPV) and 91% sensitivity. CNNs' performance on image data is almost unchallenged with respect to other machine learning techniques [14, 15, 31]. That said, CNNs are black boxes, and even with advancements in model interpretability, it is still hard for humans to fully understand the decision process of these models [26]. This problem is exacerbated by the need for task-specific and clinically oriented interpretability solutions and a lack of regulations in AI research and development. Therefore, interpretability and decision explainability remain open challenges.

Another body of related machine learning literature uses structured (i.e., tabular) data to assist with pandemic planning and decision making. Brinati et al. [3] showed the feasibility of using machine learning with routine blood test data to achieve similar performance metrics as the gold standard of RT-PCR tests in discriminating between negative and positive COVID cases. Machine learning can also assist in forecasting the severity of the infection for those patients already diagnosed with COVID [12, 18, 29, 32]. Machine learning for prognosis can be used at different stages of COVID infection for patients. Gao et al. [12] used blood test data, comorbidity history, and other basic patient information to assess mortality risk using a voting ensemble of supervised classifiers and achieved high AUC (area under the receiver operating characteristic curve) scores ($> 90\%$) across several patient cohorts. Assaf et al. [1] used supervised classification and information at the time of hospital admission to predict the risk of further deterioration. With a small dataset of 162 total hospitalized patients, from which 25 cases deteriorated into critical COVID, they reported a high AUC score of 0.9 that outperformed clinical risk assessment parameters such as APACHE II score [19].

Machine learning classification tasks usually fall under the two categories of supervised and unsupervised learning. Supervised algorithms learn to map the input space onto the output space through encountering labeled data that includes instances of different possible values in the output space [7]. Unsupervised algorithms take on the challenge of learning from unlabeled data and are themselves tasked with discovering reoccurring trends in the data [2]. One the other hand, some machine learning algorithms (that may be supervised or unsupervised) are tasked with classification in severely imbalanced populations; these are called anomaly detection algorithms [4]. We implement AdaBoost [11] and XGBoost [6] for imbalance-aware supervised learning and OCSVM [28] for unsupervised anomaly detection

learning. AdaBoost and XGBoost are famous boosting algorithms that bring together many weak classifiers into a single robust classifier [27]. On the other hand, OCSVMs are well-known unsupervised anomaly detection algorithms that can model the decision boundaries of a high-dimensional distribution [8].

Model interpretability for COVID prognosis has been implemented for imaging data, though the interpretability methods used—Grad-CAM in Zhang et al. [33] and Li et al. [20], GSInquire in Wang et al. [30]—provide heatmap overlays that only indicate which parts of an image were important, but not what quality made them important. In the case of structured data, there are numerous model-based and model agnostic methods for interpretability. For example, Brinati et al. [3] trained random forests and leveraged the model-based feature importance. On the other hand, model agnostic interpretability methods are more appropriate when the importances of the same features are being studied across different models.

SHAP is currently one of the most widely adopted interpretability methods; its applications range from COVID prognosis [10, 29] to financial time-series analysis [23]. SHAP draws from game theory to explain the output of any machine learning algorithm by computing the marginal contribution of each data feature in the final output of the model per sample [21]. Notably, SHAP values provide a sample-level understanding of a model's output by calculating a set of values for each sample and allocating optimal credit to each attribute of that sample. SHAP values are usually calculated only once on the validation samples and are used to study individual feature importances. SHAP does not natively allow for calculations of feature importance across multiple cross-validations, which are often appropriate for small datasets. Additionally, SHAP is designed to assess individual feature importances, however, in our dataset, we create new features that aggregate individual features into groups by clinical interest. Thus, we implemented novel extensions to SHAP for multiple cross-validations and feature groups.

3 Data

This retrospective study included all diagnosed SARS-CoV-2 patients from a small Canadian province up until March 26, 2021, under research ethics approval. The diagnosis was confirmed with reverse-transcriptase polymerase chain reaction (RT-PCR) testing. The total number of diagnosed patients in the study was 988, where 42 (4%) cases were at-risk (resulting in hospitalization and/or death), and 8 ($< 1\%$) cases resulted in death. For each patient, information regarding their sex, age, comorbidities, and perceived symptoms at the time of diagnosis were available, totaling 27 features (Table 1). Figure 1 shows the case counts by age, while Figure 2 shows the distribution of comorbidities by age. Distributions of all other features, which include sex, mode of infection acquisition (not included in the modeling attributes), symptoms, and outcome, are shown in Figure 3. We note that the time period of the study spans before and after vaccine availability, however, the data did not contain vaccination status; thus, underlying distributions of features for patients testing positive/negative may have changed during the study, particularly with respect to age.

Table 1: Features and feature groups. All features and feature groups other than age are 0/1 values.

	Feature groups				
Feature name	Comorbid-ities	Constit-utional symptoms	Fever symptoms	Respiratory symptoms	Symptoms combined
Age					
AMI status	✓				
Asthma status	✓				
COPD status	✓				
Cough				✓	✓
Diarrhea					✓
DM status	✓				
Fever			✓	✓	✓
Feverish/Chills			✓	✓	✓
General weakness		✓			✓
Headache		✓			✓
Heart-failure status	✓				
Hypertension status	✓				
IHD status	✓				
Irritability		✓			✓
Loss of appetite		✓			✓
Loss of smell/taste					✓
Nausea		✓			✓
Pain		✓		✓	✓
Painful swallowing					✓
Runny nose					✓
Sex					
Shortness of breath				✓	✓
Small red purple spots					✓
Sore throat					✓
Stroke status	✓				
Symptomatic					✓

AMI: acute mycardial infraction; COPD: chronic obstructive pulmonary disorder; DM: diabetes mellitus; IHD: ischemic heart disease

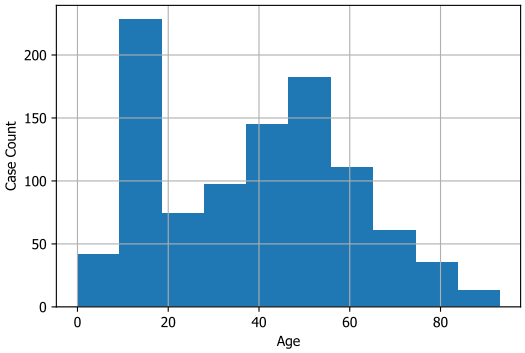

Figure 1: Distribution of patient ages

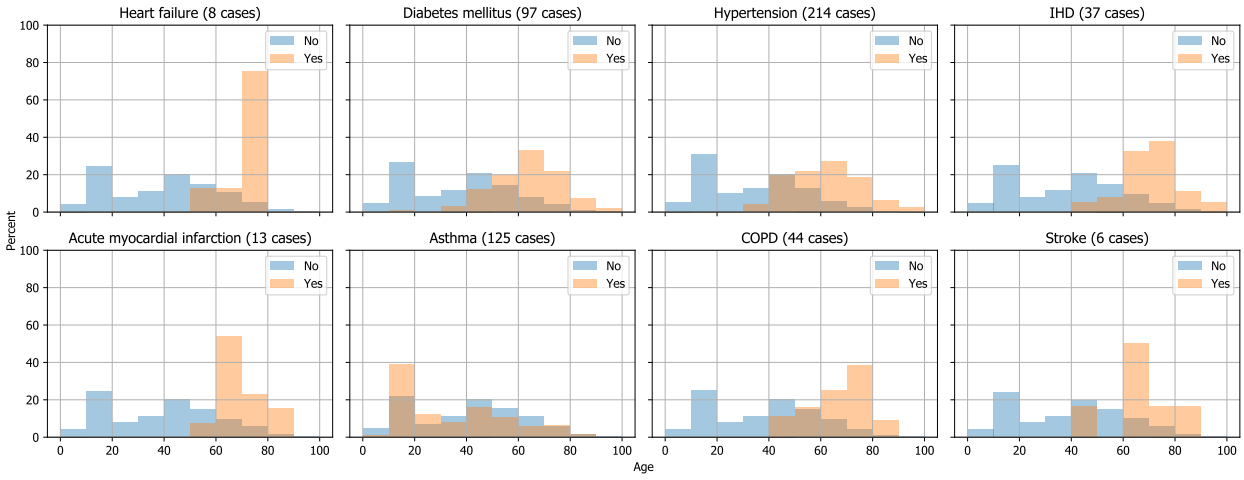

Figure 2: Density of patient comorbidities by age. Yes = at risk; No = not at risk.

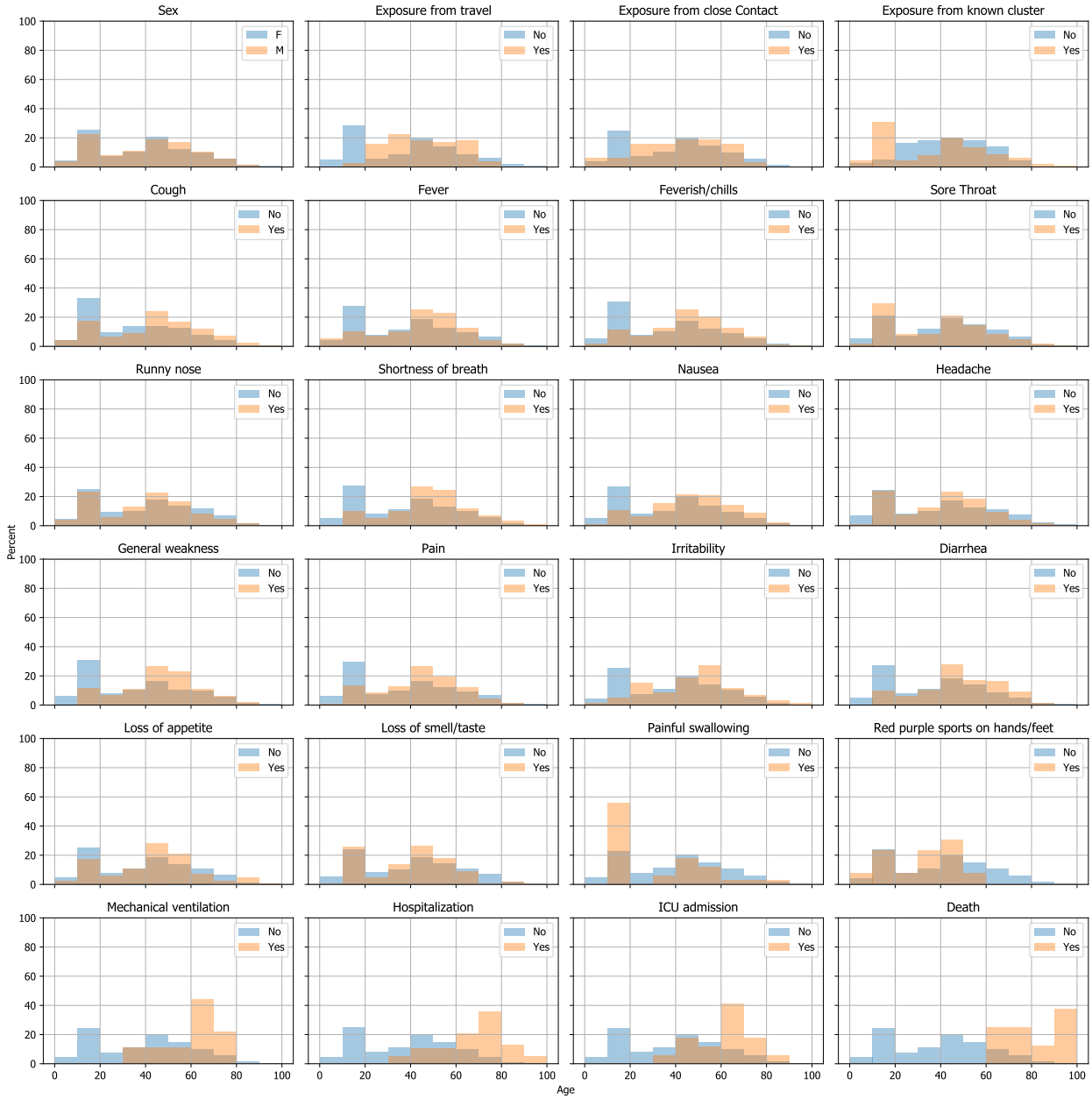

Figure 3: Density of patient sex, mode of infection acquisition, symptoms, and outcomes by age. Yes = at risk; No = not at risk.

4 Methods

We employ imbalance-aware supervised classification (AdaBoost and XGBoost) and unsupervised anomaly detection (OCSVM) algorithms to identify patients likely to experience severe COVID outcomes. Unsupervised anomaly detection is employed to detect both at-risk and deceased cases. However, the supervised models only detect the at-risk cases due to the severe class imbalance and too few records of deceased cases in the available dataset.

To gain statistical evidence regarding the models' performance stability, we implement pipelines of five times repeated 5-fold cross-validation (CV) for tuning and evaluating each model. Slightly different versions of the repeated CV are implemented for supervised and unsupervised experiments settings. Both versions are discussed in detail during the following subsections. We additionally calculate SHAP values across all models to add model explainability to our framework.

4.1 Imbalance-aware supervised classification

We implement imbalance-aware AdaBoost and XGBoost models for binary classification of patients as either normal or at-risk. The models are provided with sample instances of both outcomes during the training. To manage the class imbalance in the data, both models take advantage of custom sample weights. For AdaBoost, we tune the penalty for mislabeling the positive (at-risk) samples to 15 times more than mislabeling negative samples, while for XGBoost we tune the penalty to 300 times more.

To go beyond simple CV and address possible performance stochasticity due to the small dataset size, we repeat 5-fold CV five times with different seeds for the randomized data splitting. Each experiment consists of 25 training and evaluation iterations. The performances from all these iterations are aggregated to obtain the optimal hyperparameters and validate the results. In each iteration of the supervised settings, we use stratified splitting to train and evaluate the models on 80% and 20% portions of the entire dataset, respectively.

4.2 Unsupervised anomaly detection

We implement two unsupervised anomaly detection OCSVM models to detect the at-risk cases and deceased cases as anomalies. The models only access the normal (not at-risk, not deceased) class during the training and learn to model the distribution of the normal population. After training, the models are able to classify samples from both the normal and at-risk (or deceased) populations as either normal or out-of-distribution (anomaly). Due to requiring samples only from the majority class during the training, OCSVM is well-suited for classification with severely imbalanced data, as in our COVID dataset.

Due to the small number of samples from the anomaly populations (42 (4%) at-risk, 8 ($< 1\%$) deceased), we slightly modified the five times repeated 5-fold CV implementation. We isolate all anomaly samples at each iteration and split the data that only includes the majority samples into 80% and 20% train and test normal-only sets. The normal-only train set is used to train the models, and the normal-only test set is shuffled with all the anomaly

samples to create the final test set. Then, similar to the supervised experiments, the results from the 25 iterations are aggregated for tuning and validation.

4.3 Interpretability

To integrate SHAP into the five times repeated 5-fold CV, we calculate a set of SHAP values for the test set samples during each iteration. To be able to later compare results between different models, at the end of each CV we normalize SHAP values across all samples and all features to have mean 0 and standard deviation 1. By the end of the 25 experiments, we have calculated five different sets of SHAP values per sample in the dataset. The results are aggregated for each sample by averaging the values across the five sets. We designate this set of average SHAP values as the final SHAP value set for the samples. In collaboration with our clinical partners, we also define five feature groups of clinical interest: constitutional symptoms, respiratory symptoms, fever symptoms, comorbidities group, and symptoms group (Table 1). Note that symptomatic feature is an individual feature from the training data, which indicates to the model whether or not a patient had experienced symptoms. On the other hand, the symptoms feature group is the bundle of all symptom-related features that allows us to study the correlated contribution of all members of the group. We calculate the sum of SHAP values for the features belonging to each group as the group's marginal contribution to the output of the models.

5 Results

The mean and standard deviation of the true positive rate (TPR), false positive rate (FPR), F1 score [13], and AUC for each prediction tool are reported in Table 2 (see Appendix A for ROC curves; note that OC-SVM is not compatible with ROC curves). For AdaBoost and XGBoost, different discrimination thresholds yield different TPR/FPR results. For XGBoost, we derived the optimal discrimination thresholds by minimizing TPR − FPR. For the AdaBoost model, the default discrimination function implemented in the Python `scikit-learn` library [24] was optimal.

Figures 4 and 5 illustrate the SHAP value features importances of each feature and feature group for at-risk and deceased prediction, respectively. The SHAP importances for XGBoost

Table 2: Performance metrics. Bold: best performance for at-risk prediction

	Label	TPR	FPR	F1	AUC
XGBoost	at-risk	$\mathbf{0.96 \pm 0.05}$	0.19 ± 0.08	0.30 ± 0.1	$\mathbf{0.92 \pm 0.02}$
AdaBoost	at-risk	0.89 ± 0.11	$\mathbf{0.17 \pm 0.07}$	0.29 ± 0.11	0.89 ± 0.05
OCSVM	at-risk	0.82 ± 0.01	0.27 ± 0.03	$\mathbf{0.54 \pm 0.03}$	N/A
	deceased	0.95 ± 0.06	0.27 ± 0.03	0.22 ± 0.02	N/A

are shown here as it was the best performing predictor, however, the feature importances from the other models are generally in agreement with these importances (Appendix B), and we specifically observe that feature group importances for all methods are similar, relative to each other (Figure 6). Notably, symptoms as a group are consistently more important than comorbidities as a group. Furthermore, shortness of breath is the most important individual indicator for at-risk prediction, while being symptomatic is the most important individual indicator for death; in both predictions, the symptoms feature group is most important indicator overall. While age is an important indicator in both predictions, the relationship between high/low age and outcome is not always consistent, as evidenced by the mixed red/blue spots in the SHAP graphs.

These graphs also help validate the models' understanding of what information is essential in the early prognosis of COVID patients. For example, in agreement with the common wisdom regarding COVID, the models identified age, shortness of breath, fever, and diabetes as the most significant individual risk indicators. Note that contrary to the AdaBoost and XGBoost models, higher feature values are associated with a negative impact on the model output for the OCSVM models due to the OCSVM implementation in `scikit-learn` which labels anomaly predictions as -1, while AdaBoost and XGBoost predict the at-risk labels as 1.

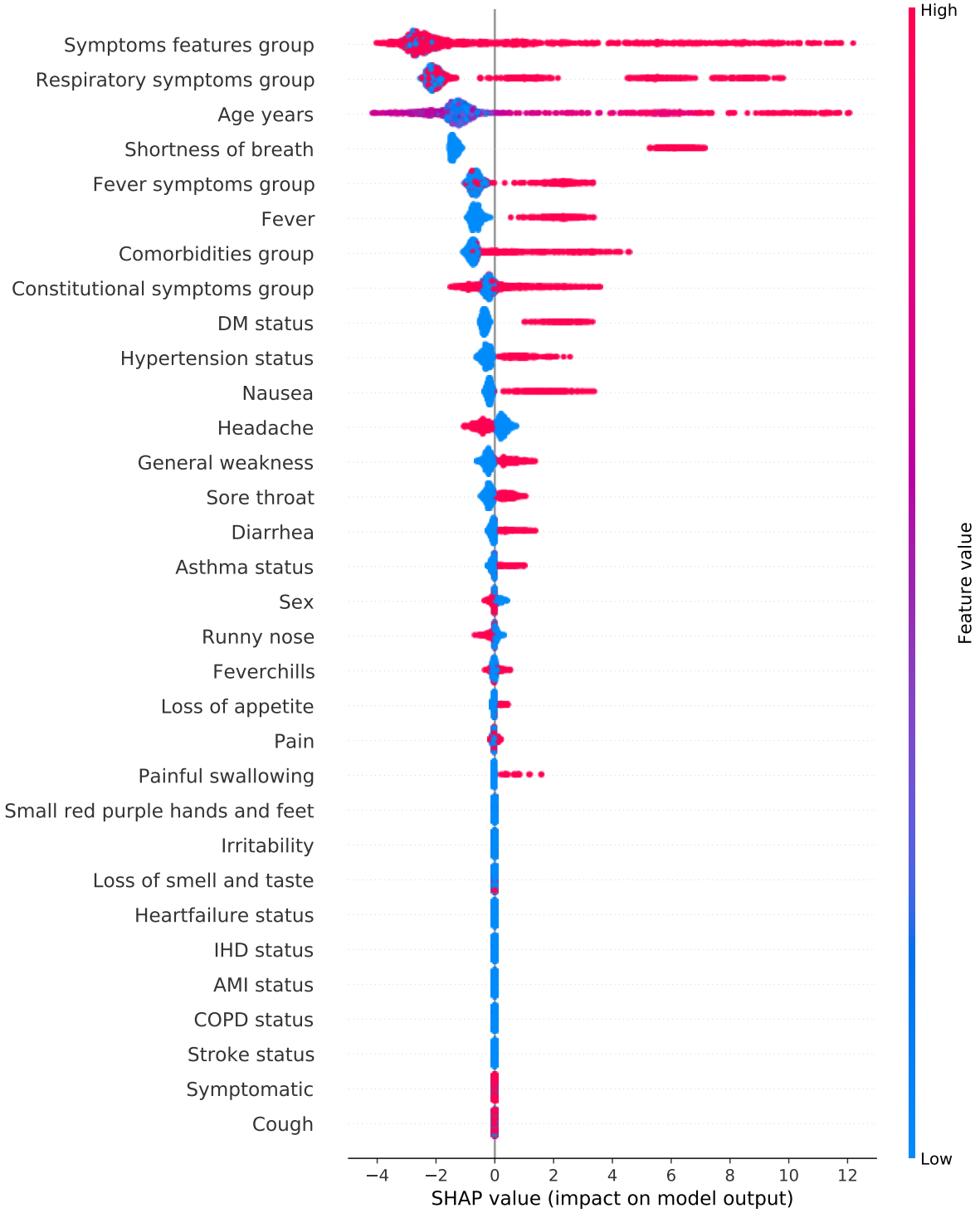

Figure 4: Scaled SHAP feature importance for XGBoost at-risk prediction

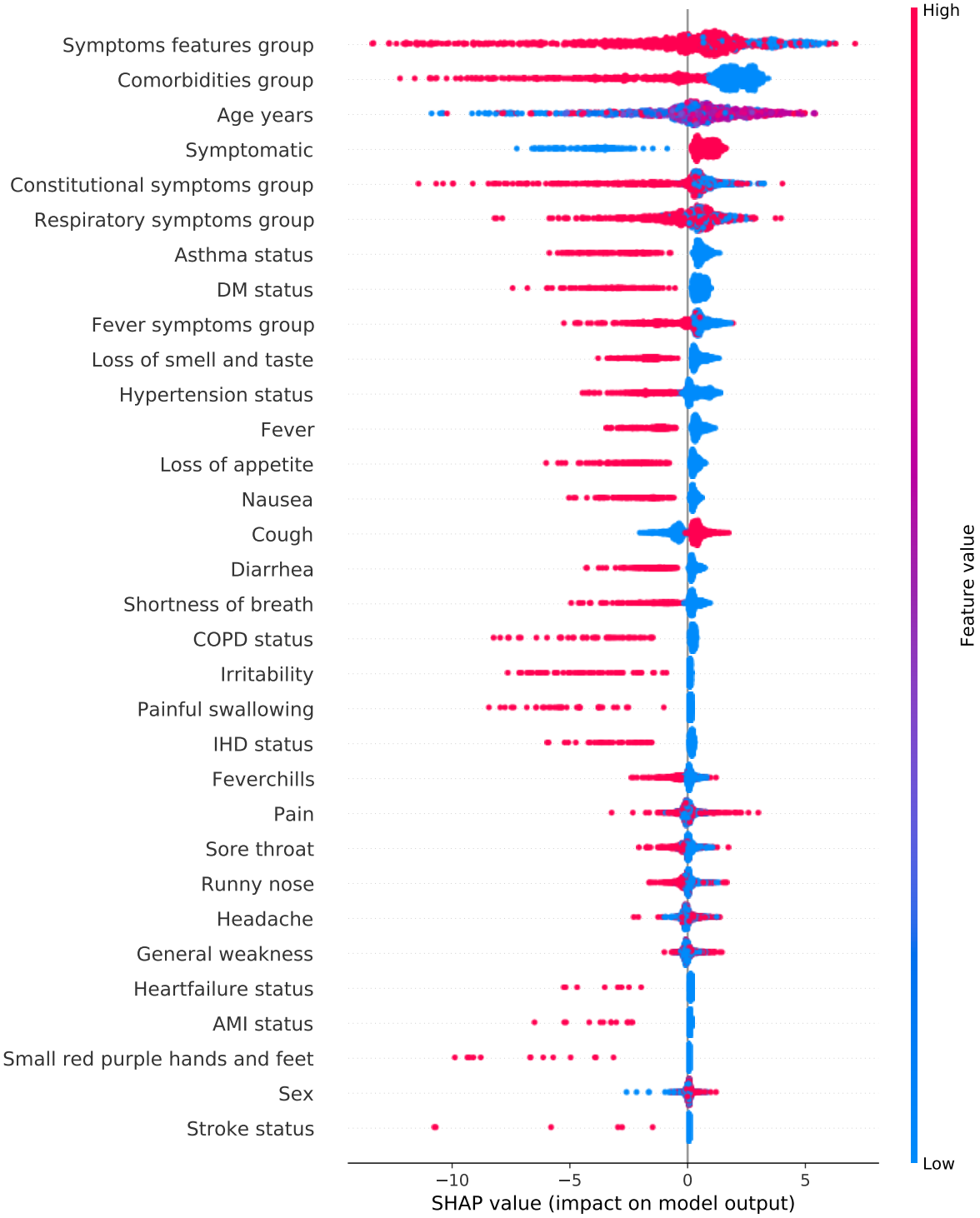

Figure 5: Scaled SHAP feature importance for OCSVM deceased prediction

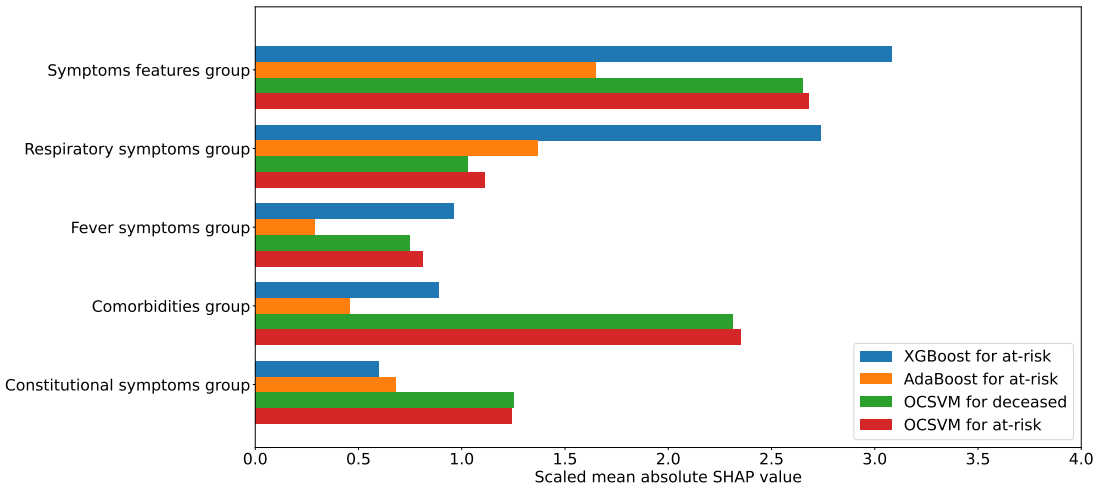

Figure 6: SHAP feature group importance for all predictors, after scaling SHAP values for all features to [0,1] within each predictor for visual clarity

6 Discussion

The novelty of our study lies in combining transparency with significant predictive power despite minimal data, and is leveraged to make the first COVID patient outcome predictions at the time of RT-PCR testing. This work provides a generalizable early prognosis and rapid triaging framework that can be quickly adopted and integrated into the testing process to inform advanced resource allocation decisions in the event of a sudden and widespread disease outbreak similar to COVID.

Our approach does not depend on data that may be time-consuming or expensive to collect; in fact, our data can almost entirely self-reported by the patient, with the exception of fever status, which can be quickly collected during testing. Most machine learning literature on COVID prognosis takes advantage of accurate results from expensive diagnostic tools such as radiography imaging [20, 25, 30] or blood tests [1, 3, 29]. Such data is cumbersome to gather for those infected with SARS-CoV-2, and it may not be collected on a mass scale by default; further, this data can only be collected once the patient is in the hospital, at which point, advanced resource allocation is no longer possible. Instead, our framework puts little to no added stress on the system because it depends only on attributes such as age, minimal comorbidities background, and apparent symptoms at the time of diagnosis. Therefore, the framework is easily integratable into the current healthcare infrastructure at regional or national levels.

Further, we have shown significant predictive power, on par with the best performances among the literature, even with a small dataset of less than 1000 COVID patients. For at-risk prediction, we obtained an AUC of 92% with XGBoost, compared to 86% [29] and 92% [1], obtained with 18000 and 6995 patient records, respectively. Fewer studies train models specifically to predict mortality due to COVID. Subudhi et al. [29], Yan et al. [32] leverage

feature-rich and large datasets to compare the performance of several supervised models for detecting COVID mortality from hospitalized patients and report high AUC measures (93% and 99%). Even though this performance is higher than our OCSVM, one should note that the performances are not fully comparable since our model is trained on a smaller dataset, for which only anomaly detection is an appropriate methodology, and predicts mortality at the time of the RT-PCR diagnosis, much earlier than hospitalization.

Additionally, we are using machine learning tools to go beyond mere predictions. On the micro-level, high performance and sample-level model explainability through SHAP provide clinical insight for each patient. On the macro-level, the aggregated sample-based explanations can guide assessment of risk factors in a local population, which provides further ability to allocate scarce resources appropriately.

The advantage of our framework lies in that the predictions are made at the time of RT-PCR positivity, and only a minimal number of such events are required for model training. In addition, because the framework is effective with only a small dataset, it can be applied to small regions to capture specific characteristics of the regional population and to accommodate changes regarding viral transmission factors and/or vaccination. However, it is important to note that the risk factors identified here are relevant only to the time period, population, and outcomes of the current study, and we again emphasize the lack of vaccination status data. For clinical implementation, a number of revisions must be taken into consideration, including collection of vaccination status, and possibly which vaccines were taken. Changes in the new virus variants or vaccination status may result in different constellations of symptoms and different degrees of severe infection rates. Therefore, the framework requires continuously updated data. Additionally, the occurrence of events such as hospitalization or ICU admission to some extent depends on the capacity and willingness of the local health care system to provide specialized care to patients.

7 Conclusion

We validated and explained the behavior of three machine learning models to predict, at the time of RT-PCR testing, whether COVID patients are at risk of severe outcome and/or death. This early prediction allows for advanced allocation of scarce healthcare resources, including ventilators and clinical staff. Our models were trained on minimal data from a small cohort of fewer than 1000 patients. The models are easy to use, can be adjusted based on regional needs, and showed significant and consistent predictive power in the prognosis of patients at risk. Finally, we incorporated our extended SHAP explainability into the framework to gain insights from the models' behavior, allowing for clinical validation of their decisions. A future avenue of study is to validate the results and study the differences when applying the framework on multiple cohorts of geographically separated patients who may have significantly different underlying risks. Additionally, the models should be regularly updated with new cases, should contain vaccination status for improved accuracy, and their predictions should be compared with clinical predictions.

References

[1] Dan Assaf, Ya'ara Gutman, Yair Neuman, Gad Segal, Sharon Amit, Shiraz Gefen-Halevi, Noya Shilo, Avi Epstein, Ronit Mor-Cohen, Asaf Biber, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and Emergency Medicine*, 15(8):1435–1443, 2020.

[2] Horace B Barlow. Unsupervised learning. *Neural Computation*, 1(3):295–311, 1989.

[3] Davide Brinati, Andrea Campagner, Davide Ferrari, Massimo Locatelli, Giuseppe Banfi, and Federico Cabitza. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. *Journal of Medical Systems*, 44(8):1–12, 2020.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

[6] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[7] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine Learning Techniques for Multimedia*, pages 21–49. Springer, 2008.

[8] Shifei Ding, Zhibin Zhu, and Xiekai Zhang. An overview on semi-supervised support vector machine. *Neural Computing and Applications*, 28(5):969–978, 2017.

[9] Juan L Domínguez-Olmedo, Álvaro Gragera-Martínez, Jacinto Mata, and Victoria Pachón Álvarez. Machine learning applied to clinical laboratory data in spain for COVID-19 outcome prediction: Model development and validation. *Journal of Medical Internet Research*, 23(4):e26211, 2021.

[10] Fernando Timoteo Fernandes, Tiago Almeida de Oliveira, Cristiane Esteves Teixeira, Andre Filipe de Moraes Batista, Gabriel Dalla Costa, and Alexandre Dias Porto Chiavegatto Filho. A multipurpose machine learning approach to predict COVID-19 negative prognosis in são paulo, brazil. *Scientific Reports*, 11(1):1–7, 2021.

[11] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[12] Yue Gao, Guang-Yao Cai, Wei Fang, Hua-Yi Li, Si-Yuan Wang, Lingxi Chen, Yang Yu, Dan Liu, Sen Xu, Peng-Fei Cui, et al. Machine learning based early warning system

enables accurate mortality risk prediction for COVID-19. *Nature Communications*, 11 (1):1–10, 2020.

[13] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer, 2005.

[14] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[15] Sakshi Indolia, Anil Kumar Goswami, Surya Prakesh Mishra, and Pooja Asopa. Conceptual understanding of convolutional neural network-a deep learning approach. *Procedia Computer Science*, 132:679–688, 2018.

[16] Atiqa Khalid and Sana Ali. COVID-19 and its challenges for the healthcare system in pakistan. *Asian Bioethics Review*, 12(4):551–564, 2020.

[17] Abolfazl Zargari Khuzani, Morteza Heidari, and S Ali Shariati. COVID-classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images. *Scientific Reports*, 11(1):1–6, 2021.

[18] Hyung-Jun Kim, Deokjae Han, Jeong-Han Kim, Daehyun Kim, Beomman Ha, Woong Seog, Yeon-Kyeng Lee, Dosang Lim, Sung Ok Hong, Mi-Jin Park, et al. An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: Retrospective cohort study. *Journal of Medical Internet Research*, 22(11):e24225, 2020.

[19] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache II: a severity of disease classification system. *Critical Care Medicine*, 13(10): 818–829, 1985.

[20] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest ct. *Radiology*, 2020.

[21] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[22] Ian F Miller, Alexander D Becker, Bryan T Grenfell, and C Jessica E Metcalf. Disease and healthcare burden of COVID-19 in the united states. *Nature Medicine*, 26(8): 1212–1217, 2020.

[23] Karim El Mokhtari, Ben Peachey Higdon, and Ayşe Başar. Interpreting financial time series with shap values. In *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pages 166–172, 2019.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Jawad Rasheed, Alaa Ali Hameed, Chawki Djeddi, Akhtar Jamil, and Fadi Al-Turjman. A machine learning-based framework for diagnosis of COVID-19 from chest x-ray images. *Interdisciplinary Sciences: Computational Life Sciences*, 13(1):103–117, 2021.

[26] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020.

[27] Robert E Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[28] Bernhard Schölkopf, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt, et al. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588. Citeseer, 1999.

[29] Sonu Subudhi, Ashish Verma, Ankit B Patel, C Corey Hardin, Melin J Khandekar, Hang Lee, Dustin McEvoy, Triantafyllos Stylianopoulos, Lance L Munn, Sayon Dutta, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ Digital Medicine*, 4(1):1–7, 2021.

[30] Linda Wang, Zhong Qiu Lin, and Alexander Wong. COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12, 2020.

[31] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights Into Imaging*, 9(4):611–629, 2018.

[32] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 2(5):283–288, 2020.

[33] Jianpeng Zhang, Yutong Xie, Guansong Pang, Zhibin Liao, Johan Verjans, Wenxing Li, Zongji Sun, Jian He, Yi Li, Chunhua Shen, et al. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. *IEEE Transactions on Medical Imaging*, 2020.

A ROC curves

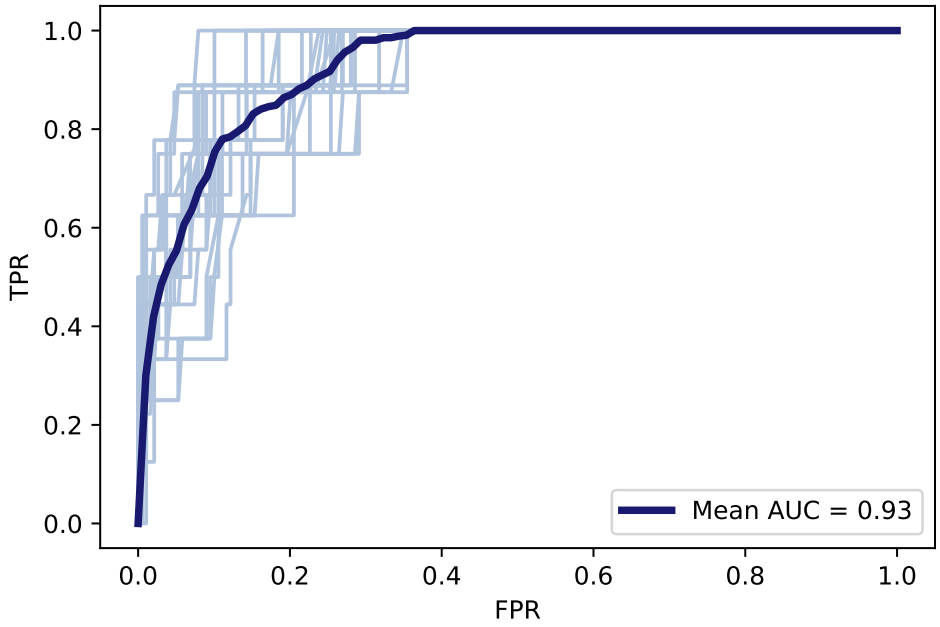

Figure 7: ROC curve for XGBoost at-risk prediction across all validation iterations; bold line is the average curve

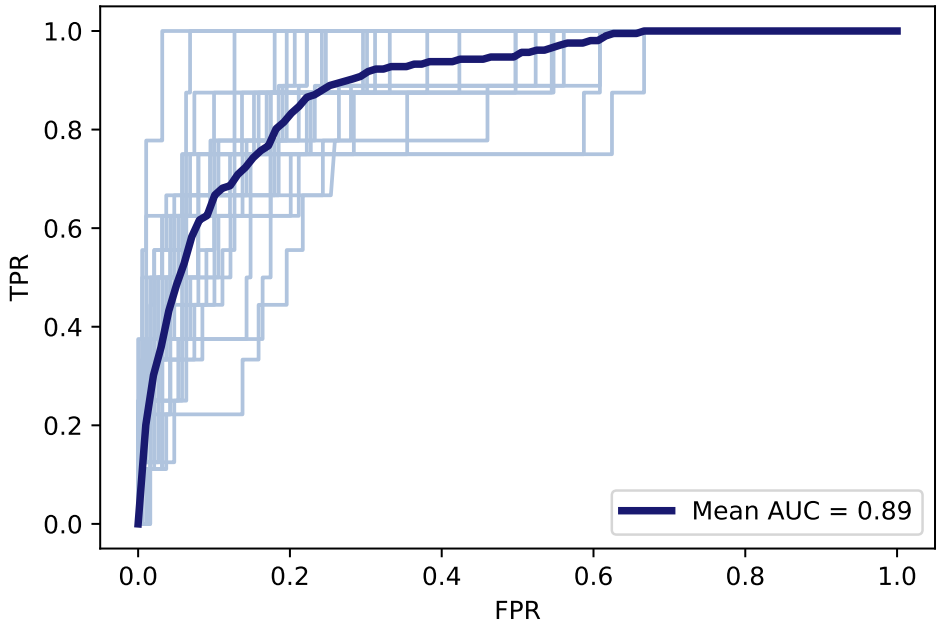

Figure 8: ROC curve for AdaBoost at-risk prediction across all validation iterations; bold line is the average curve

B SHAP values for all methods

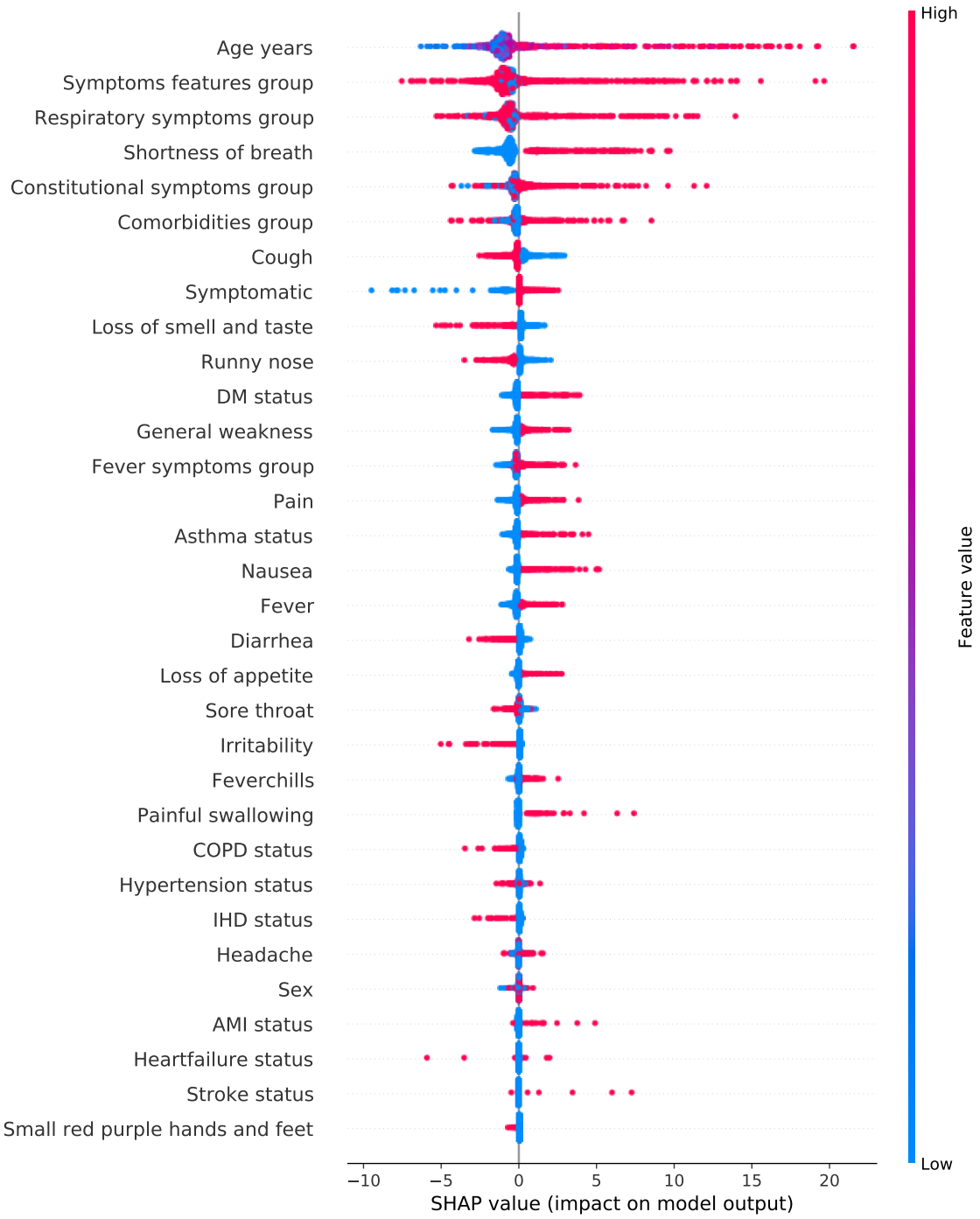

Figure 9: Scaled SHAP feature importance for AdaBoost at-risk prediction

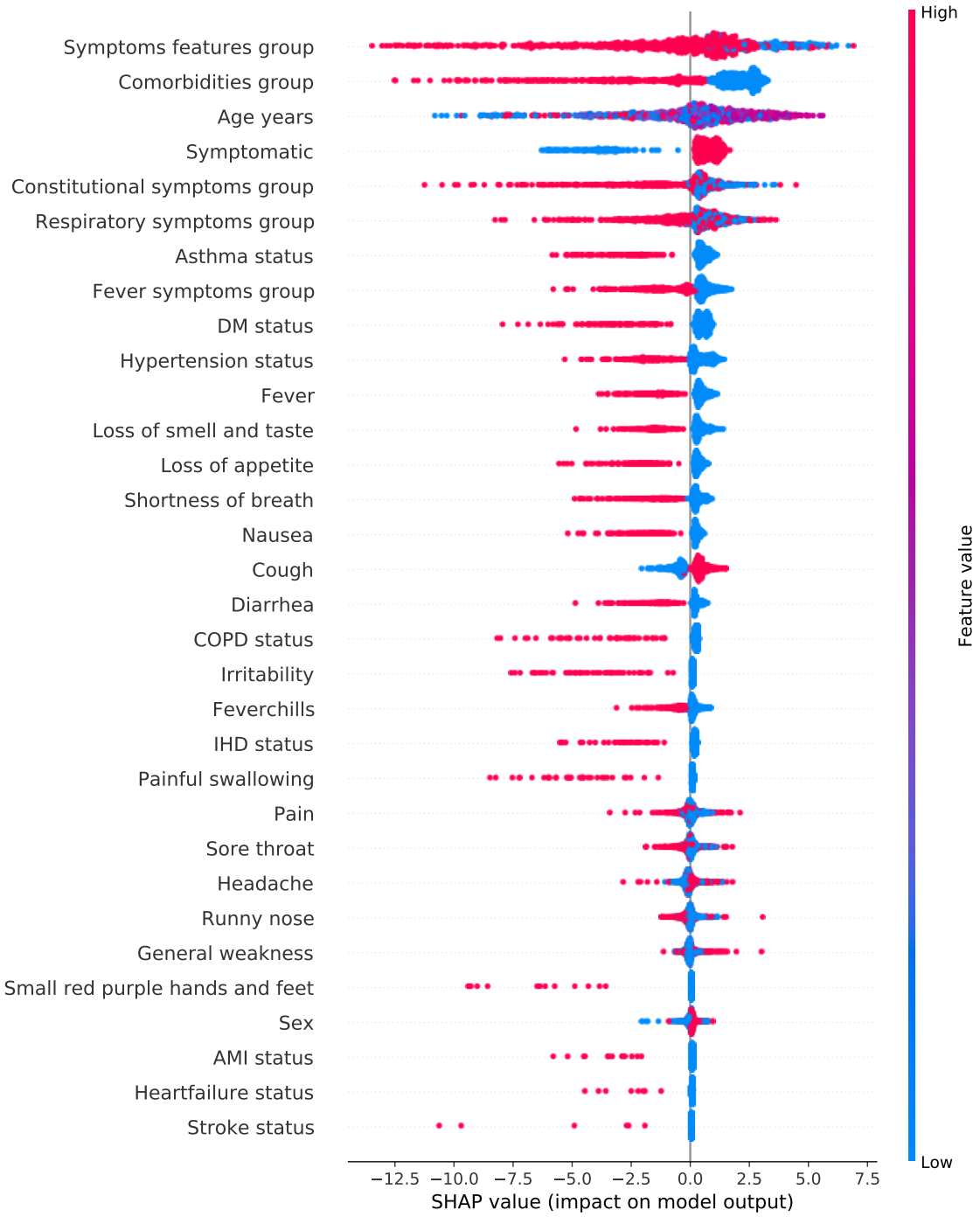

Figure 10: Scaled SHAP feature importance for OCSVM at-risk prediction

C Optimized hyperparameter values

Table 3: Optimized hyperparameters; bold is the optimal hyperparameter found.

Model	Optimized hyperparameters
XGBoost	`eta`:range(0.5, 3, 0.1); selected: **2.6** `max_depth`: range(1,11,1); selected: **3** `gamma`: [**0**, 5] `min_child_weight`: [**1**, 30] `max_delta_step`: range(0, 6, 1); selected: **2** `subsample`: [0.5, **1**] `sampling_method`: [**uniform**, gradient_based] `alpha`: [**0**, 3] `tree_method`: [**auto**, hist, approx, exact] `scale_pos_weight`: range(1, 30, 1); selected: **23** `refresh_leaf`: [0, 1] `grow_policy`: [**depthwise**, lossguide] `num_round`: range(10, 41, 10); selected: **30** normal samples' weight: **0.01** at-risk samples' weight: range(1, 10, 1); selected: **3**
AdaBoost	`n_estimators`: [100, **200**, 300] `algorithm`: [SAMME, **SAMME.R**] `learning_rate`: [0.05, 0.1, **1**] normal samples' weight: **1** at-risk samples' weight: range(10, 50, 1); selected: **30**
OCSVM	`gamma`: [scaled, **auto**] `nu`: range(0, 1, 0.1); selected: **0.2** `kernel`: [linear, poly, **rbf**]

XGBoost implementation from Python XGBoost library [5]; OCSVM and AdaBoost implementation from [24]