Machine learning models to predict the diagnosis and risk of COVID-19 from clinical data in Switzerland

Elia Escoffier, Sélène Ledain, Micol Bassanini
EPFL, CS-433, Machine Learning

Abstract—**The COVID-19 pandemic quickly saturated limited healthcare resources such as diagnostic tests and hospital beds, and has highlighted the need for decentralised management. This study develops and compares several machine learning predictive models for the diagnosis and risk stratification of COVID-19 using various intelligently selected combinations of resources. Using a cohort of 170 patients presenting suspicions of pneumonia at a Swiss outpatient department, we train logistic regression, random forest and neural network classifiers to discriminate COVID+ vs COVID- (diagnosis) and, for COVID+ patients (n=88), whether they require hospitalisation or not (risk stratification). We compare models derived with several combinations of features that are selected using decision trees, recursive feature elimination, forward feature selection, Pearson's correlation coefficient, Chi-2 correlation and LightGMB. Such models may help decentralise decision making and better allocate scarce resources.**

I. Introduction

The COVID-19 pandemic has saturated healthcare resources across the globe and highlighted the need for more efficient probabilistic triage. Since symptoms are non-specific, it is difficult to predict diagnosis and prognosis. This can lead to misallocation of limited resources such as diagnostic kits and hospital beds. While RT-PCR and computed tomography (CT) have become gold standard diagnostic and prognostic tools, they are costly centralised resources. For diagnosis, RT-PCR sometimes had delays of 7 days rendering the results epidemiologically and clinically worthless. Early identification high risk patients is associated with lower mortality [1]. Additionally, these tests suffer from significant false positives [2] and there is thus a need to create a composite reference standard that may improve performance. Improving probabilistic triage will allow us to identify patients with the highest need for these limited resources. For risk stratification, CT imaging is an expensive technique and exposes patients to radiation. Lung ultrasound (LUS) has potential as a cheaper, quicker and easy-to-use alternative to CT imaging for risk stratification in COVID-19 patients [3]. We explore the predictive value of a range of clinical features to replace the high-resource tests of RT-PCR and CT.

II. Aim & Objectives

The aim is to develop and compare predictive models for the diagnosis and prognosis of COVID-19 using several combinations of clinical features. The objectives are to

- preprocess the datasets to minimize bias from missingness and confounding.
- identify the most predictive features in order to produce a parsimonious and pragmatic model for clinical use
- derive a set of machine learning models to predict the diagnosis and prognosis of COVID-19
- compare performance when using combination of features representing cheap and expensive resources.
- explore whether machine learning can improve prognostication performance for expert labels of lung ultrasound (LUS).

III. Methods

A. Study Design and dataset

In this retrospective cohort study, 170 adult patients were recruited at the CHUV emergency department during the first wave of the COVID epidemic in Switzerland (from March 6 to April 3 2020). Inclusion criteria was a suspicion of pneumonia and informed consent.

Features: For all models, only features that were collected on day 0 were used so that the models could be used to make predictions at triage. The number of features for each model was limited to 10 for pragmatic clinical use and the methods for feature selection are described below. *Clinical features:* All patients completed a clinical questionnaire, clinical exams and blood tests, resulting in 205 different features. Features are split into low (lowR) and high (highR) resource categories, where low resource features are those that can be gathered in a questionnaire by telephone contact. High resources features include those where patient contact and invasive/centralised procedures (X-ray/CT/blood tests) are required. *Lung ultrasound:* Additionally, a lung ultrasound (LUS) was performed on all COVID+ patients (and thus only available for the prognostic model). Acquisition was performed at 8 thoracic sites and images were labelled by experts as one of 4 ordinal pathological groups (0 for normal, 1 for b-lines pattern, 2 for confluent b-line, 3 for irregular pleural line, 4 for consolidation) as previously described [3]. LUS is considered in combination with clinical data and alone and compared to existing risk stratification models [3].

Outcomes: *Diagnosis:* the binary COVID+/- class was defined by RT-PCR on a nasopharneal swab on all 170 patients. *Prognosis:* Prognosis was only performed on the 88 COVID positive patients who were followed up for 30 days during which time they were labeled with a 3-class ordinal multiclass outcome feature: outpatients, hospitalized, intubated/death. Two binary classifications are also proposed: (A) requiring or not hospitalisation, and (B) mild dyspnea from those who require intubation or die.

B. Data pre-processing

*1) **Reducing collinear dimensionality**:* To observe the correlation and redundancy between features, we created 5 interaction features as a product of pairs of variables to summarize information. From a medical point of view, some features represent the same information or are scores combining existing parameters. 42 features were removed because of redundancy.

*2) **Handling missing values**:* There were $\frac{13514}{34850}$ (40%) missing data points across the 205 x 170 matrix. We established a procedure to remove them, starting with analyzing both the number of missing values per patient and per feature. A patient is removed when its row contains more than 70% missing features (n=2). A feature column is removed when >50% of values are missing. Features with at most 20% missingness were imputed with a *K*-nearest neighbors algorithm (using the mean value of the 5 nearest neighbors) to stay as consistent as possible. Features with missingness >20% that were interesting to keep nonetheless were further investigated for bias. Here, we computed the dependence of feature missingness with the outcome group to determine whether the features was Missing Completely At Random (MCAR), Missing At Random (MAR, missingness is random, but can be fully explained by other variables with complete information) or Missing Not At Random (MNAR, the value of the variable that's missing is related to the reason it's missing i.e. biased relative to the outcome). If features were MCAR (or "independent"), they could be kept and imputed. MNAR features were removed due to the risk of bias. The LUS dataset contained 58 features corresponding to scores for the different thoracic sites. Specific variables indicating if the observation was missing were removed.

*3) **Feature selection**:* The goal is to minimize the number of variables in the different models to under 10, in order to ensure that the model is pragmatic for clinicians to use at testing time. The first technique applied is a Lasso regularization : it selects certain features as it forces certain coefficients to be set to 0 by restricting the sum of the absolute value of the regression coefficients to be less than a fixed value. Their importance ranking allows interpretability. Lasso regularization was completed with Pearson's correlation (which kept a selected number of features mostly correlated with the output), a similar method using a Chi-2 test, tree based algorithms such as LightGBM (a gradient boosting framework that uses decision tree learning) or feature importance in Random Forests, recursive feature elimination (RFE) and forward feature selection. The feature selection between the different models was compared and performances assessed with logistic regression as a baseline.

C. Predictive models

As a baseline, a **logistic regression (LR)** model is implemented on the 2 feature sets (low and high resources).

Random Forest (RF): The hyperparameters were adjusted with 4 fold cross validation and grid-search on a range of values for the number of trees, maximum amount of features for splitting a node, maximum tree depth, minimum number of data points placed in a node before splitting, number of data points allowed in a node and sampling data points with or without replacement.

Neural Network (NN): The architecture of the NN was defined by comparing Random Search and Grid Search. These were used for neural architecture search (NAS). The search space is parametrized by the number of layers and the number of neurons for each layer, restricted to NNs with at most 4 hidden layers and 40 neurons per layer. For reasons of computational cost, the analysis regarding the Random Search was performed on a validation set instead of cross validation as in Grid search. Both methods had as choosing criterion the accuracy on the validation set. To stabilize the learning process and mitigate overfitting, a probability dropout layer that randomly excluded 50% of the layer units and their connection at each training step.

IV. Results

A. Diagnosis

By taking in consideration all the low resource (lowR) features (n= 42), the logistic regression (LR) model predicted correctly on average 72% of the time. When using all the high resource (highR) features (n=71), LR had a 9% improved mean accuracy of 81%.

As shown on Figure 1, lasso regularization identified $\frac{18}{42}$ and $\frac{30}{71}$ features as non-significant within the lowR and highR feature groups respectively, which were subsequently removed.

By applying the 6 feature selection methods described above, the most relevant (selected >3 times among the methods) features were kept. It decreased the lowR features to 7: *fever, duration of fever, presence of sputum, age, history of chronic obstructive pulmonary disease (COPD) and its severity, and history of chronic inflammatory disease* (see Table I) and the highR to *fever and fever duration, age, presence of sputum, history of COPD and its severity, leukocyte count and radiological infiltrate (on chest X-ray)* (see Table I).

*1) **Random Forest**:* A total of $\frac{8}{38}$ lowR (*fever, age, symptom duration, cough, sputum, history of chronic kidney disease, history of COPD and COPD decompensation in the last 6 months*) and $\frac{7}{68}$ highR features (*age, history of COPD,*

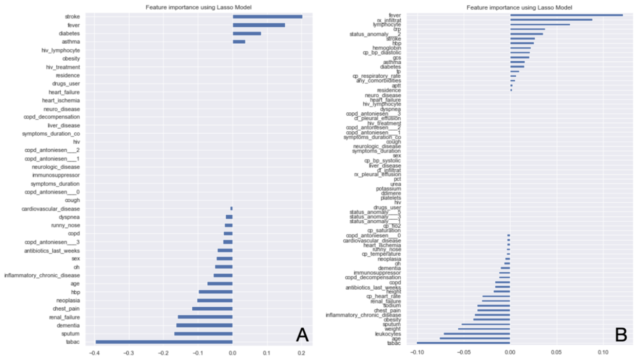

Figure 1: Coefficients for the different features, using a lasso regularization. A : lowR diagnosis model. B : highR diagnosis model.

COPD decompensation in the last 6 months, history of heart failure or heart ischemia, fever, diastolic blood pressure, respiratory rate and leukocyte count) were retained using decision trees and RFE respectively. The optimized lowR model obtained an AUROC of 80% with a recall of 74% (1100 trees, with at most 60 levels, take up to the square root of number of features for splitting a node and take at least 2 data samples for splitting a node). The highR model has 1000 decision trees with at most 20 levels, a minimum of 5 data points in a node, a minimum of 15 data points considered before splitting a node and no replacement when sampling data points. This model obtains an AUROC of 86% and 77% recall (see Table I).

*2) **Neural Network**:* The selected lowR and highR features were the same as those selected for LR. The lowR trained model with 2 hidden layers and 8 neurons per layer obtains an area under the ROC of 72% and 74% recall (see Table I). For the highR model, the AUROC was 80%. The NN has the best recall compared to LR and RF, with 81% (see Table I).

	LR		RF		NN	
	(LowR)	(HighR)	(LowR)	(HighR)	(LowR)	(HighR)
N. of features	5	6	8	7	5	6
Accuracy	0.72	0.80	0.76	0.82	0.71	0.80
Recall	0.72	0.80	0.74	0.77	0.74	0.81
F1	0.72	0.80	0.79	0.84	0.73	0.80
Precision	0.74	0.81	0.85	0.92	0.75	0.79
AUROC	0.72	0.79	0.80	0.86	0.72	0.80

Table I: Model for diagnosis. Green label : best model LowR. Yellow label : best model HighR.

In short, the Random Forest (RF) performed best on both low and high resource (lowR and highR) features with an AUROC of 80% and 86% respectively. LR and NN were 8% worse for diagnosis using lowR features and and 6% worse when using the highR feature set (see Table II).

B. Prognosis

Results for the multiclass prognostic classification achieved systematically only around 50% accuracy likely due to the small sample size in each class. Two prognosis models were therefore developed in the larger binary classes: **A = can go home VS need additional care at hospital** and **B = need for respiratory assistance VS need to go in care unit because of serious respiratory problems.**

For the prognosis A, using both lowR and highR features, LR had 85% accuracy. The predictions with LR on combined lowR and highR data for the prognosis B have 69% correct classifications. After comparing the features most frequently selected by the 6 different selection methods, those retained are *presence of pulmonary infiltrate, sputum, procalcitonin, sex, age, and heart rate* for prognosis A and *history of stroke, age, blood urea, diabetes, leukocyte count, respiratory rate and oxygen saturation* for the prognosis B classification.

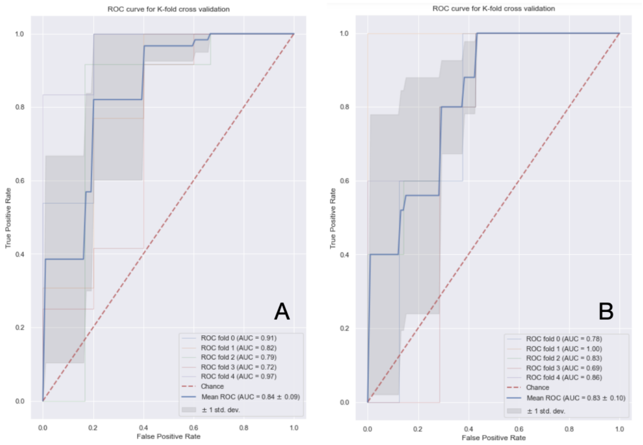

Figure 2: ROC curves using a 5-fold stratified cross validation with a logistic regression in the case of prognosis. A : hospitalisation vs no hospitalisation. B : mild dyspnea vs intubation/death.

*1) **Random Forest**:* For the prognosis A classification (looking at need of hospitalisation or not), using RF on the clinical data (combining lowR and highR) gives 78% accuracy. The features selected with Pearson's correlation are: *oxygen saturation, fraction of inspired oxygen, c-reactive protein, heart rate, leukocyte count, respiratory rate, presence of pulmonary infiltrate*.

When a RF is implemented for the prognosis B comparing dyspnea vs intubation/death, with features selected using decision trees *(leukocytes, c-reactive protein, respiratory rate, total protein, fraction of inspired oxygen, procalcitonin)*, the accuracy is of 78% (see Table II).

*2) **Neural Network**:* The NN took the same clinical features as selected for the LR. In prognosis A there is 78% accuracy and in the prognosis B classification there is 79%, which is the best performance among the various models for prognosis B on clinical data (see Table II).

C. Prognosis with LUS

Again, multi-class classification didn't perform well, with accuracy of around 60% (Figure 3).

Thus, models for binary prognosis were also developed using LUS alone and combination with clinical data. Only the binary outcomes (prognosis A and B) are presented here.

Best results were obtained using Random Forest on only LUS (see Table II). Adding LUS data to the clinical data did not always improve the results: for prognosis A accuracy increases accuracy by 1% and for prognosis B it actually drops by 7% compared to using only clinical data. The **Random Forest performs best with only LUS features**. For the prognosis A, with 6 features selected after RFE *(qld_1, qaig_0, qlg_4, qpid_3, qpig_3, qpig_4)* there is 96% accuracy. For the prognosis B classification, the model (using *qld_1, qaig_0, qlg_4, qpsg_1, qpig_3, qpig_4* selected after RFE) has 92% accuracy.

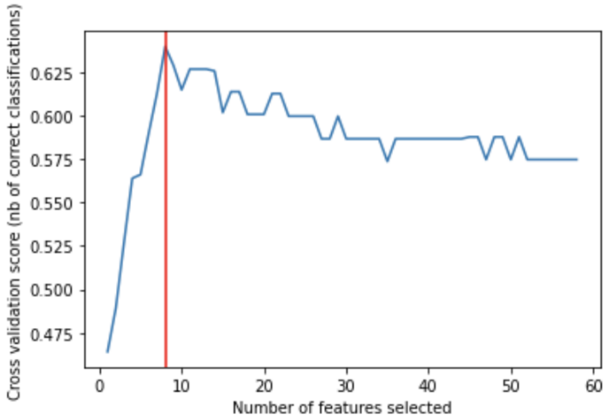

Figure 3: Estimation of the optimal number of features and corresponding accuracy, using a 7-fold cross-validation, performed on LUS data only, using a multiclass classification.

		LR		RF		NN	
		(A)	(B)	(A)	(B)	(A)	(B)
Clinical	N.feat	6	7	8	7	6	6
	Acc.	0.83	0.76	0.78	0.78	0.78	0.79
LUS	N.feat	3	3	6	6	3	3
	Acc	0.80	0.73	0.96	0.92	0.80	0.77
Clinical	N.feat	4	4	8	8	4	4
+LUS	Acc	0.88	0.75	0.79	0.71	0.89	0.70

Table II: Model for prognosis. (A) : can go home vs need additional care at hospital. (B) : need for respiratory assistance vs need to go in care unit. Green label : best model using clinical data. Yellow label : best model with LUS data. Red label : best model with both sets.

In short, the Random Forest (RF) performed best in both prognosis classifications A and B (home vs need care at hospital and need for respiratory assistance vs need to go in care unit) (see Table II).

V. Discussion

Diagnosis: Using both the low and high resource feature sets, the **random forest clearly outperforms** logistic regression and neural network in terms of AUROC (see green and yellow labels in Table I for lowR and highR respectively). In medical diagnosis, this is expected to reduce false positives and false negatives. Feature selection reveals that ***fever* is the strongest predictor of COVID-19** among all variables from the clinical set (see Figure 1).

Prognosis: Prognosis is more difficult to predict than diagnosis due to the unbalanced representation of each classes in the dataset. Changing the multi-classification problem into several binary classifications helped improve the results. The Figure 2 illustrates that a reasonable trade-off between recall and specificity exists in the 2 final models proposed by feature selection methods using LR concerning binary prognosis, as both ROC curves tend to get closer to the perfect classifier. Both models have at least 80% chance to detect the good prognosis outcome. The prognostic models also highlight the importance of LUS data and need for only few features. As shown on Figure 3, with a multi-class outcome a maximum of 60% accuracy can be reached with 8 features, and binary outcomes make it even better (Table II). **Prognosis using random forest on LUS data performed better for both binary tests**, and only 6 features in total are needed to discriminate all types of outcomes with a better accuracy. It shows that **LUS is a good way to reduce costs and improve efficiency in triage**. If clinical data is included, then LR or NN is more interesting.

VI. Limitations

The datasets that were provided are of limited size and representative of only a certain part of the population. To ensure that our diagnostic and prognostic tests can be deployed to a wider group, further studies are needed to train and validate our models. It is also assumed that the labels and medical tests are correct whereas it could be possible that patients were misdiagnosed [2].

Using highR models gives better results compared to the using only lowR data. Often, only 1 or 2 additional expensive tests are performed and the necessity of bringing patients in for this might be discussed. If the priority is to decentralise diagnosis and keep detection cheap, such a test might not yet be attractive.

Finally, performing LUS tests on only certain regions of the lungs may not be realistically implementable in a medical context, where patient condition and physiology might restrict examination.

VII. Future work

As COVID-19 is still virulently circulating, information about new patients, from diverse age groups and socio-economic backgrounds, can be collected and help to improve actual models. The same is true to better understand the efficiency of LUS, and would also help check if features are group specific or not, in order to make models more generalisable.

References

[1] Sun Q. Qiu H. Huang M. Lower mortality of covid-19 by early recognition and intervention. 2020.

[2] Aaron S. Kesselheim Steven Woloshin, Neeraj Patel. False negative tests for sars-cov-2 infection — challenges and implications. *The New England Journal of Medecine*, 2020.

[3] Dr Thomas Brahier Pr Jean-Yves Meuwly Dr Olivier Pantet Marie-Josée Brochu Vez Hélène Gerhard Donnet Dr Mary-Anne Hartley Pr Olivier Hugli Dr Noémie Boillat-Blanco. Lung ultrasonography for risk stratification in patients with covid-19: a prospective observational cohort study. 2020.

[4] Martin Jaggi Mary-Anne Hartley. Deepbreath: Diagnostic pattern detection for covid-19 in digital lung auscultations. 2020.

[5] Ozturk T Talo M Yildirim EA Baloglu UB Yildirim O Rajendra Acharya U. Automated detection of covid-19 cases using deep neural networks with x-ray images. 2020.

[6] Using chi-square statistic in research. https://www.statisticssolutions.com/using-chi-square-statistic-in-research/.

[7] C. Leavitt. Uncovering missing not at random data. 2019. https://towardsdatascience.com/uncovering-missing-not-at-random-data-8d2cd3eda31a.

[8] J. Brownlee. An introduction to feature selection. 2014. https://machinelearningmastery.com/an-introduction-to-feature-selection/.

[9] How to diagnose the missing data mechanism. 2013. https://www.theanalysisfactor.com/missing-data-mechanism/.

[10] J. Brownlee. knn imputation for missing values in machine learning. 2020. https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/.

[11] Bergstra J. Bengio Y. Random search for hyper-parameter optimization. 2012.

[12] Yu K. Sciuto C. Jaggi M. Musat C. Salzmann M. Evaluating the search phase of neural architecture search. 2019.

[13] J. Brownlee. Multi-class classification tutorial with the keras deep learning library. 2016. https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/.

[14] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. 1995.

[15] Understanding auc - roc curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[16] Liu X et al Wang Y, Kang H. Combination of rt-qpcr testing and clinical features for diagnosis of covid-19 facilitates management of sars-cov-2 outbreak. *J Med Virol*, 2020.