COVID-19 Predictions using Machine Learning

Authors
Laure Font, Laurine Lang, Annina Stuber
Supervisor
Marie-Anne Hartley, Machine Learning and Optimization Laboratory / intelligent Global Health
CS-433 Machine Learning, EPFL Lausanne, Switzerland

***Abstract*—Combining engineering, selection of medical features, and machine learning, we implemented binary classifiers to predict the diagnosis and prognostic outcome of COVID-19. Different models were used based on resources split according to cost and complexity. For both diagnostic and prognostic models, the selection of more costly features did not significantly increase classifier performance, and therefore does not justify the need for expensive diagnostic tests when basic over-the-phone questions or auscultation suffice. A model was also implemented to compare our predictions to a study where the scoring of lung ultrasounds by specialists was used to predict the prognostication of patients [1], and comparable results were obtained.**

I. Background

The COVID-19 disease, caused by SARS-CoV-2 virus, was first identified in Wuhan (China) in December 2019 and was declared a pandemic by the World Health Organization in March 2020. The health systems are overwhelmed by the number of patients requiring medical care and the rapid propagation of the disease. Currently as no cure exists for COVID-19 and as an early intervention could reduce the mortality [2], there is a need to improve diagnostic and prognostic predictions in order to better allocate resources where they are needed most. While high-accuracy tests do exist (RT-PCR and CT scans), these rely on expensive and limited centralised resources, causing bottlenecks and delays that render results useless, and are unaffordable in low-income settings. Thus, there is a need to find predictive patterns in clinical data and exams that may assist triage in peripheral screening centers. However, due to the non-specific presentation, the predictive patterns in these data are not easily discernable by humans. This study explores machine learning methods to optimize these predictions based on heterogeneous data.

II. Aim and objectives

Aim: To improve the probabilistic methods to predict the diagnosis and prognosis of COVID-19 while conserving resources using machine learning.

Objectives:

- To develop and compare predictive models for the diagnosis of COVID-19 based on several selected feature sets (free, cheap, expensive)
- To develop and compare predictive models for prognostication using several selected feature sets as above.
- To develop a prognostic model using machine learning methods on human experts analyses of Lung Ultrasound (LUS) and compare it to existing scores.

III. Data exploration and feature engineering

A. Data

The dataset includes records of 170 adult patients who presented at the emergency department of the Lausanne University Hospital (CHUV) with lower respiratory tract infections between March 6th and April 3rd 2020, among which 88 tested positive for SARS-CoV-2 [1]. A total of 303 features were recorded as part of the health assessment. The following subcategories englobe features based on the difficulty of obtaining the information:

- "**Free**" (i.e. possible to perform via telephonic contact): demographic data (*sex, age, weight, height*), medical history (*arterial hypertension, diabetes, asthma, renal failure...*), symptoms (*symptoms duration, dyspnea, fever...*);
- "**Cheap**" (i.e. requiring a clinical exam): signs (*auscultation, vital signs*), point of care tests such as Lung Ultrasound (LUS) with expert analyses of 10 different zones of the lungs;
- "**Expensive**" (i.e. invasive and costing significant centralised resources): blood tests (*Angiotensin 2, Interleukin-6, Interleukin-8, TREM1 and C-Reactive Protein*) and centralised paraclinical exams (*chest X-ray*).

In the following sections, cheap features will refer to the combination of both free and cheap features; and expensive features will refer to the combination of free, cheap and expensive features. The outcomes investigated were:

- **Diagnosis** as measured by nasopharyngeal RT-PCR (88 patients COVID+ and 82 patients COVID-)
- **Prognosis** as assessed with a 30-day follow-up where 38 patients were categorised as mild (outpatient or hospitalised without oxygenotherapy) and 50 patients were categorised as severe (hospitalised with oxygenotherapy, the Intensive Care Unit (ICU) or fatal).

For diagnosis, all samples were used. Because the prognosis is related to the severity of COVID-19, only the 88 COVID-19 positive patients were retained. The prognosis prediction was intended to be multi-class to discriminate risk on an ordinal scale from outpatient, hospitalized, ICU and death. Yet, due to the limited number of patients in each class, low performance was obtained and the above binary label was preferred.

B. Study Design

First, features regarded as irrelevant for the study were discarded (e.g. features collected after the outcome had occurred,

administrative features, etc.). The remaining 181 features were divided into 3 feature sets described above (free, cheap, expensive). Additionally, LUS features were considered alone. As the diagnostic classifier is intended to be used in various environments, features with environmental or geographic bias towards certain societies, countries or cultures (sex, age, weight, height, smoking habit, alcohol consumption and drug abuse) were discarded. These features were, however, kept for the prognostic classifiers, as they may influence the disease severity. As an example: in some societies, an elderly person may be more likely to be diagnosed with COVID-19 (nursing homes in Switzerland) while this would not be the case in other cultures. However, it has been observed that an elderly person, regardless of geographical location, is more likely to have more severe symptoms than a young person [3].

C. Data exploration

During an extensive data exploration, histograms were displayed and statistical analyses were conducted.

Relationships between features were represented with Jaccardian similarity matrices and with Pearson correlation matrices. In order to reduce the dimensionality of the data, features which we considered as combined information were preferred to small sub-categories, redundant or strongly correlated features. For example, *neurologic disease* was retained to represent *stroke* and/or *dementia* and/or *neuro disease*; *cardiovascular disease* was kept to represent *heart ischemia* and/or *heart failure*; *symptoms duration* was preferred to correlated *fever duration* and *cough duration*; diastolic and systolic blood pressures were kept and the average, *tam*, removed.

Binary features with rare occurrences (less than 5% of non-null entries) were considered to lead to biased predictions, and therefore discarded as well.

D. Feature engineering

For clarity, some features from the blood tests were categorized into different levels of gravity with respect to the difference to the average physiological range (*Hemoglobin, Leukocytes, Neutrophils, Lymphocyte, Platelets, Urea, Asat, Alat, Total bilirubin, CRP, Sodium, Potassium*) [4], [5].

Chest X-ray and LUS were only performed on COVID-positive patients, and were thus solely considered for prognosis prediction. For each zone in which LUS was performed, five boolean features representing the different observations were recorded: 'NA' (no measurement), 0 (normal), 1 (pathologic B lines), 2 (confluent B lines), 3 (thickening of the pleura) and 4 (consolidations). To reduce dimensionality, a single feature was created for each zone combining the observations as mentioned above using integers (0-4); 0 was also attributed to missing recording to avoid a bias toward not recorded zones. Additional binary features were created to indicate whether the LUS abnormalities were multifocal (abnormalities reported in at least 2 zones) or bilateral (abnormalities affecting both lungs). A total LUS score and a normalized LUS score (nLUS), normalized by the number of measured zones (corrected for missing values), were also evaluated.

E. Handling missing data

Before the imputation of missing values, we identified whether the missing values were biased towards a certain label (not missing at random: NMAR) or if they were unbiased and thus missing at random (MAR). Each feature was binarized (0: missing, 1: otherwise). A Chi Square test of independence was performed between each binarized feature and the label (*COVID-19* for diagnosis, *outcome* for prognosis). Features with a p-value smaller than 0.05 were considered NMAR, and thus removed. If the test assumptions were not met (at least 5 samples per category in the contingency table), a Fisher's exact test was performed instead. Features with a p-value smaller than 0.05 were discarded. Eventually, 3 expensive features were NMAR for the diagnosis datasets; 1 free and 14 expensive features were NMAR for the prognosis datasets.

Various methods were considered to process missing values. Due to the limited number of samples, it was impossible to simply discard samples with missing data. However, features with more than 40% of missing values were discarded. K-nearest neighbours (k-NN), a method commonly used to impute missing data, was considered for the remaining features. Because k-NN does not perform well in high dimensions, especially with few data points, Pearson correlation was computed between all features to impute the missing values of each feature using only the features with which it was mildly correlated. However, with a threshold of 0.5, most features in the free and cheap datasets were only correlated with themselves or with a single feature. k-NN was thus not considered. Due to small datasets and insufficient correlation, missing values were simply replaced by the median of each feature. The mean was not considered because it is less robust to outliers in skewed distributions.

IV. Methods

Before applying any machine learning algorithm, each dataset was split into a train set and a test set, by setting aside 20% of the data. The train sets were then normalized for each feature by subtracting the min and dividing by *max-min*. The parameters of the train sets were used to normalize the test sets.

The same algorithms were considered to predict the diagnosis and prognosis: Ridge Regression, Logistic Regression (using L1-, L2- or no regularization), Random Forest and Neural Network. Adaptive and Gradient Boosting were also considered but were quickly set aside due to their low performance. Grid search was carried out with a 5-fold cross validation on the training set to determine the best hyperparameters of each algorithm: the strength of regularization α for Ridge Regression and λ for Logistic Regression, the number of trees and the depth of each tree for Random Forest, the number of nodes and the number of hidden layers for Neural Network.

Feature selection was an important aspect of the pipeline, due to the large number of features and small amount of samples. It was also fundamental to use the least possible amount of clinical signs to save resources when diagnosing the disease or when predicting its outcome. Different approaches

were explored to reduce the dimensionality of the previously tuned models. Random Forest, regularized Logistic Regression and Ridge Regression are embedded methods which naturally assign each feature a coefficient related to its importance in the classification process. The best features were selected by setting a threshold of 0.02 for the minimum feature importance (SelectFromModel). Wrapper methods (Recursive Features Elimination and Backward Features Selection) were alternatively considered. In Backward Feature Selection, features performances were evaluated either by the p-value of each feature (ordinary least squares, OLS) or by the performance of the model when the feature was removed (Backward). For Neural Network, Local Interpretable Model-Agnostic Explanations (LIME), returning the contribution of each feature for the prediction of a single sample, was used to extract the most important features for each model. The maximum number of selected features was set to 12 for pragmatic clinical reasons. The hyperparameters of the algorithms were then further fine-tuned on the reduced datasets.

Different metrics were computed to estimate the performance of each model: accuracy, sensitivity (recall), specificity and Area Under the Receiver Operator Curve (AUROC). Sensitivity and AUROC are of particular interest for model and feature selection, as false negatives are more important to avoid in comparison to false positives in the context of infectious diseases. However, sensitivity maximizes the detection of patients with COVID-19 (for diagnostic predictions) and with severe outcome (for prognostic predictions), at the expense of the other classes. The classifiers thus tend to increase the sensitivity at the expense of the precision. As such, AUROC was preferentially used. In cases where multiple models had close AUROC scores, the other metrics; accuracy, sensitivity and specificity were considered in descending order. If the decision was still ambiguous, models with features of lesser cost were preferred. Classification results on the test sets were displayed in the form of confusion matrices.

V. Results

A. Objective 1: Diagnosis of COVID-19

The best models for predicting the diagnosis of COVID-19 are summarized in Table I. Based on AUROC, Random Forest (with 102 trees of max. depth 3) was selected as being the best approach for diagnostic prediction with free features (AUROC 0.742). For the same diagnostic prediction using cheap and expensive features, unregularized Logistic Regression had the highest performance with an AUROC 0.768 and 0.775 respectively. Importance of significant features for each model can be found in Figure 1a, darker shade represents higher importance of the feature for the model.

B. Objective 2: Prognosis of COVID-19

The same considerations as in section V-A were taken into account when selecting the best classifiers for prognostication. The best models and feature selection methods can be found in Table I: unregularized Logistic Regression for the free features (AUROC 0.833), Random Forest for the cheap (AUROC 0.875) and expensive (AUROC 0.958) features (with respectively 402 trees of max. depth 3 and 202 trees of max. depth 3). The importance of the retained features can be found in Figure 1b.

C. Objective 3: Lung ultrasound based prognostication

Unregularized Logistic Regression with Backward selection resulted in the model with the highest AUROC score on the LUS dataset (AUROC 0.783) (see Table I). The most relevant features for classification were the nLUS score, the LUS score, the multifocality and the bilaterality.

Diagnosis					
Method	Dataset	Accuracy	Sensitivity	Specificity	AUROC
Random Forest (SelectFromModel)	Free	0.735	0.8	0.645	0.742
Logistic (OLS)	Cheap	0.765	0.8	0.737	0.768
Logistic (OLS)	Expensive	0.765	0.867	0.684	0.775
Prognosis					
Method	Dataset	Accuracy	Sensitivity	Specificity	AUROC
Logistic (OLS)	Free	0.889	1	0.667	0.833
Random Forest (Recursive)	Cheap	0.889	0.917	0.833	0.875
Random Forest (SelectFromModel)	Expensive	0.944	0.917	1	0.958
Logistic (Backward)	LUS	0.812	0.9	0.667	0.783

TABLE I: *Evaluation of performance on the test sets for the best COVID-19 diagnostic and prognostic predictors using free, cheap, expensive or LUS features*

VI. Discussion

A. Objective 1: Diagnosis of COVID-19

Comparing the three best models with their respective subsets of features, we can see that they have rather similar scores for each metric (Table I). Looking closer at expensive and cheap models, we can state that the costs at which expensive features would come, may not sufficiently increase the performance of the classification to justify the expenses. The similar performances of these two models can be backed up by the fact that they share 8 of their 12 most significant features (see Figure 1a). The free model relies heavily on the information from the symptoms duration and the presence of fever, which both either play no, or only a small role in the two other models. Interestingly, medical history of renal failure is observed as a predictive factor for a COVID-19 positive test. It is unexpected as Acute Kidney injury rather than chronic renal failure has been identified as a predictor of COVID-19, as it has been stated as a common complication [6]. Although not observed in our model as the feature was discarded in the prognosis datasets due to its rare occurrences (see Section III), renal failure is listed by the CDC as a risk factor for severe COVID-19. Chronic kidney disease is indeed known to create immunodeficiencies that may predispose infection [7].

B. Objective 2: Prognosis of COVID-19

Predicting prognosis (in need of a hospital bed or not) had seemingly more successful results than diagnostic prediction.

All three models have a relatively high AUROC score (Table I), the other metrics follow the same trend. Although the expensive model has the highest overall metrics, all three models categorize the patients' needs quite loyally. In Figure 1b we can observe less overlap between the most relevant features of each model than in the diagnosis models (see section VI-A). Solely among the free features, age played a major role in deciding if a patient should be hospitalized, the duration of the symptoms came in as a strong second, corroborating the general current knowledge on the disease progression [8], [3]. When features requiring a clinical exam and invasive tests are considered, the FIO_2 % (fraction of inspired oxygen) in the blood appeared as an important indicator of the patient's status. This measure has since been used to monitor whether the condition of a person is deteriorating and more medical care would be needed [9]. However, as FIO_2 % is solely different from the ambient oxygen content for patients receiving oxygen and thus hospitalized, it is correlated with the outcome and biases the model. *cp_respiratory_rate*, the number of breaths per minute, also appeared as an important feature. This feature coincides with the need for more oxygen or with difficulty breathing, and thus with our current knowledge on the need of more intense care (oxygen or respirator).

feature	free	cheap	exp
symptoms_duration	0.22	0	0
fever	0.193	0.056	0.021
renal_failure	0.149	0.116	0
sputum	0.106	0.045	0.027
copd_antoniesen___3	0.091	0.076	0.039
cardiovascular_disease	0.059	0	0
neoplasia	0.04	0	0
neurologic_disease	0.034	0	0
runny_nose	0.033	0	0
chest_pain	0.029	0	0
antibiotics_last_weeks	0.024	0.039	0
diabetes	0.022	0	0
gcs	0	0.197	0.124
cp_heart_rate	0	0.14	0.071
cp_bp_diastolic	0	0.122	0.077
status_anomaly___3	0	0.093	0.122
dyspnea	0	0.044	0
status_anomaly___2	0	0.037	0.031
any_comorbidities	0	0.034	0
TREM-1	0	0	0.151
crp_routine	0	0	0.139
leukocytes	0	0	0.101
potassium	0	0	0.098

(a) diagnostic models

feature	free	cheap	exp
age	0.194	0.048	0
symptoms_duration	0.15	0.044	0
fever	0.112	0	0
oh	0.084	0	0
any_comorbidities	0.081	0	0
dyspnea	0.081	0	0
bmi	0.074	0.051	0
sex	0.073	0	0
hbp	0.044	0	0
neurologic_disease	0.039	0	0
chest_pain	0.038	0	0
sputum	0.031	0	0
cp_fio2	0	0.26	0.237
cp_respiratory_rate	0	0.168	0.182
cp_saturation	0	0.091	0.065
LUS_score_normalized	0	0.077	0.054
cp_temperature	0	0.076	0.039
cp_heart_rate	0	0.067	0.051
cp_bp_diastolic	0	0.056	0
LUS_score	0	0.035	0
qpig	0	0.027	0
TREM-1	0	0	0.082
creatine	0	0	0.075
IL-6	0	0	0.072
crp_routine	0	0	0.05
IL-8	0	0	0.049
glucose	0	0	0.044

(b) prognostic models

Fig. 1: *Relative importance of the most important features for the best models for the diagnosis and prognosis prediction*

C. Objective 3: Lung ultrasound based prognostication

According to Brahier et al. [1], the following metrics were reported when medical professionals looked solely at the nLUS score to distinguish between outpatients and admitted patients: AUROC 0.8, sensitivity 81%, specificity 59%, positive predictive value (PPV) 88% and negative predictive value (NPV) 45%. Since there were a total of 63 admitted patients and 17 outpatients, this resulted in an overall accuracy of 75%. PPV accounts for the number of accurate positive predictions in respect to the total number of positive predictions, NPV accounts for the correct negative predictions with respect to the total negative predictions. Our prediction of prognostics based on LUS resulted in a slightly lower AUROC of 0.783. However, all other metrics were higher (see Figure I). A PPV of 82% and a NPV of 80% were computed using the confusion matrix representing classification of the test samples. Although validation is limited to a small test set, we can conclude that our model is an improvement compared to the model based on nLUS scores. It should also be noticed that the model established by Brahier et al. was not validated on a separate set of patients, and might thus be overfitting the data.

D. Limitations

The main limitation of this study is the small sample size. By dividing the data into train and test sets, we are left with relatively few samples to train each model without overfitting the data and to validate them. This is especially the case for the prognostic data as only the 88 COVID-19 positive patients were retained. We noticed that our diagnostic models perform slightly poorer than our prognostic models on their larger validation test sets, which is unexpected as more samples were considered to train the models. This could point to the fact that we may be overfitting our models regardless of our attempts (feature restriction, cross-validation) to avoid this phenomenon. Neural Networks yielded relatively good performances before the feature selection, but after selecting only the twelve best features using LIME, the overall performance for prognosis decreased. It is likely that the selected features were not sufficient to explain the model, therefore further decreasing the performances. Having more data could allow better predictions for prognostication with Neural Network. Lastly, this cohort was recruited based on inclusion criteria of "suspicion of pneumonia" which made their clinical criteria particularly homogeneous, thus diluting the predictive potential of individual features of a more general population.

E. Future work

This pandemic has further pointed out the necessity for hospitals to share data in an anonymized way. There have been many attempts to making machine learning models for both prognostication and diagnostic predictions, however none seem to be completely satisfying [10]. Data collected individually in each research center are not sufficient to accurately train the models.

It would also be interesting to look into the misclassified individuals, to identify if they were in fact suffering from another condition which led to false classification.

REFERENCES

[1] T. Brahier, J.-Y. Meuwly, O. Pantet, M.-J. Brochu Vez, H. Gerhard Donnet, M.-A. Hartley, O. Hugli, and N. Boillat-Blanco, "Lung ultrasonography for risk stratification in patients with COVID-19: a prospective observational cohort study." [Online]. Available: https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa1408/5907893

[2] Q. Sun, H. Qiu, M. Huang, and Y. Yang, "Lower mortality of COVID-19 by early recognition and intervention: experience from jiangsu province," vol. 10, no. 1, p. 33. [Online]. Available: https://doi.org/10.1186/s13613-020-00650-2

[3] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson, "Estimates of the severity of coronavirus disease 2019: a model-based analysis," vol. 20, no. 6, pp. 669–677. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1473309920302437

[4] "Guidance for industry: Toxicity grading scale for healthy adult and adolescent volunteers enrolled in preventive vaccine clinical trials," p. 10.

[5] "Common terminology criteria for adverse events (CTCAE)," p. 155.

[6] M. K. Nadim, L. G. Forni, R. L. Mehta, M. J. Connor, K. D. Liu, M. Ostermann, T. Rimmelé, A. Zarbock, S. Bell, A. Bihorac, V. Cantaluppi, E. Hoste, F. Husain-Syed, M. J. Germain, S. L. Goldstein, S. Gupta, M. Joannidis, K. Kashani, J. L. Koyner, M. Legrand, N. Lumlertgul, S. Mohan, N. Pannu, Z. Peng, X. L. Perez-Fernandez, P. Pickkers, J. Prowle, T. Reis, N. Srisawat, A. Tolwani, A. Vijayan, G. Villa, L. Yang, C. Ronco, and J. A. Kellum, "COVID-19-associated acute kidney injury: consensus report of the 25th acute disease quality initiative (ADQI) workgroup." [Online]. Available: http://www.nature.com/articles/s41581-020-00356-5

[7] J. Tecklenborg, D. Clayton, S. Siebert, and S. Coley, "The role of the immune system in kidney disease," *Clinical & Experimental Immunology*, vol. 192, no. 2, pp. 142–150, 2018.

[8] M. W. Tenforde, S. S. Kim, C. J. Lindsell, E. Billig Rose, N. I. Shapiro, D. C. Files, K. W. Gibbs, H. L. Erickson, J. S. Steingrub, H. A. Smithline, M. N. Gong, M. S. Aboodi, M. C. Exline, D. J. Henning, J. G. Wilson, A. Khan, N. Qadir, S. M. Brown, I. D. Peltan, T. W. Rice, D. N. Hager, A. A. Ginde, W. B. Stubblefield, M. M. Patel, W. H. Self, L. R. Feldstein, K. W. Hart, R. McClellan, L. Dorough, N. Dzuris, E. P. Griggs, A. M. Kassem, P. L. Marcet, C. E. Ogokeh, C. N. Sciarratta, A. Siddula, E. R. Smith, and M. J. Wu, "Symptom duration and risk factors for delayed return to usual health among outpatients with COVID-19 in a multistate health care systems network — united states, march–june 2020," vol. 69, no. 30, pp. 993–998. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7392393/

[9] S. Carballo, T. Agoritsas, P. Darbellay Farhoumand, O. Grosgurin, C. Marti, M. Nendaz, J. Serratrice, J. Stirnemann, and J.-L. Reny, "COVID-19: Réorganisation sous toutes ses formes dans un hôpital universitaire." [Online]. Available: https://doi.emh.ch/fms.2020.08551

[10] L. Wynants, "A journey through the disorderly world of diagnostic and prognostic models for covid-19," NeurIPS 2020.