POSA310

ARTIFICIAL INTELLIGENCE APPLIED ON ADMINISTRATIVE BIG DATA TO PREDICT THE SEVERITY OF SARS-COV-2 INFECTION

BACKGROUND

SARS-CoV-2 is the viral strain responsible for COVID-19 pandemic disease. Infected patients manifest flu-like symptoms and, in the most severe cases, acute respiratory distress syndrome.

The aim of the study was to identify the main prognostic factors underlaying the severity of SARS-CoV-2 infection, using a machine learning approach.

METHODS

The analysis was elaborated using the data contained in **administrative databases** of a sample of Italian Entities involved in the project. Patients who were **hospitalized with COVID-19 diagnosis** (ICD-9 078.89) after 1st January 2020 were included into the dataset together with 13 relevant features representing age, sex and clinical history of each patient. The final dataset contains **13,364 records for 11,753 patients divided into 2 classes**:

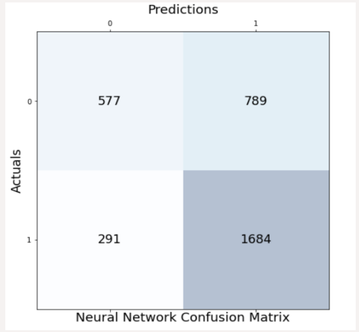
1. **Class 0: not severe symptomatology** (5,464 records)
2. **Class 1: severe symptomatology (respiratory failure /death)** (7,900 records)

A Neural Network and a Random Forest were generated using respectively Keras and Scikit-learn, two open-source library for the generation of models based on machine learning. Each model was trained on 75% of the dataset and tested on the remaining 25% to estimate the performance.

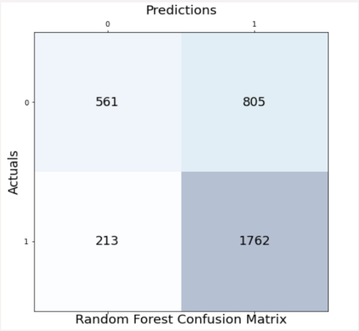
DATA PRIVACY STATEMENT. The integration of administrative datasets makes it possible to represent the patient's entire clinical history and not just individual prescriptions. The analyses were conducted on exclusively anonymous data in full compliance with privacy regulations. Clicon s.r.l. has obtained the approval as per legislation by all the Ethics Committees to analyse these data. The results are exclusively in aggregated form and never attributable to a single institution, department, doctor, individual, or individual prescribing behaviours. The study was conducted in full compliance with current legislation for retrospective studies.

RESULTS

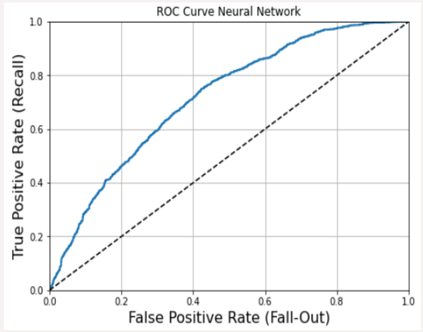
(Figure 1).



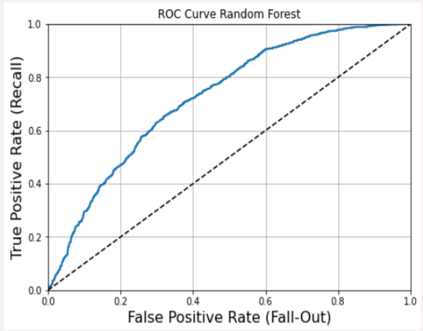
(Figure 2).



(Figure 3).



(Figure 4).



Even if the false positive rate remains high, many of the samples could be patients who would have easily developed respiratory failure but treated with the proper preventive therapies to avoid it.

MODELS PERFORMANCE

Neural Network performance on testing set:

- Accuracy: 67,7%
- Precision: 68,1%
- Recall: 85,3%
- F1 Score: 0,757
- AUC Score: 71,7%

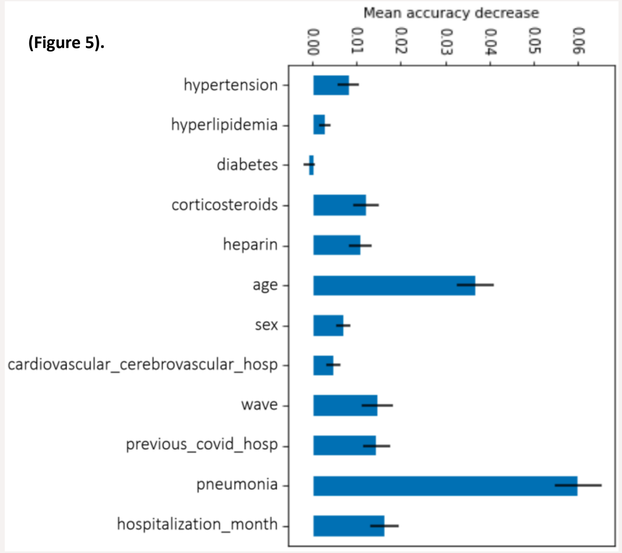
Random Forest performance on testing set:

- Accuracy: 69,5%
- Precision: 68,6%
- Recall: 89,2%
- F1 Score: 0,776
- AUC Score: 72,5%

The confusion matrices, that compares predictions made by the Neural Network (**Figure 1**) and Random Forest (**Figure 2**) with the actual classes, show that the Neural Network performs better in identfying patients belonging to class 0, while Random Forest works better in predicting pateints belonging to class 1. For both the models the AUC score was obtained measuring the respective area below the ROC curves (**Figure 3 and 4**). Generally, the models behave very well, with the Random Forest that was able to correcly classify almost 90% of severe patients in the testing set.

PROGNOSTIC FACTORS EVALUATION

(Figure 5).



The importance of the features was defined from the Random Forest model using permutation_importance, a tool included in sklearn.inspection. The tool measures the average decrease in the accuracy of the model following the removal of the considered variable (**Figure 5**). **Development of pneumonia** and **age** are **the most important prognostic factors related to the disease severity**. Further statistical analysis, on prognostic factors indicated by the model, **highlighted a higher risk in the male population and in patients with hypertension, inflammatory diseases and/or hyperlipidemia**. Finally, analysis of the periods of hospitalization shows **a significant increase in severe respect to non-severe cases during the first and second COVID waves in Italy**, probably due to the congestion of intensive care units and delays in the treatments necessary to prevent the development of respiratory failure.

REFERENCES

1. https://keras.io/getting-started/faq/
2. https://scikit-learn.org/stable/about.html

AUTHORS

Iacolare B[1], Perrone V[1], Sangiorgi D[1], Ghigi A[1], Nappi C[1], Paoli D[1], Ancona DD[2], Andretta M[3], Barbieri A[4], Bartolini F[5], Cavaliere A[6], Ciaccia A[7], Citraro R[8], Dell'Orco S[9], Ferrante F[10], Gentile S[11], Grego S[12], Procacci C[2], Ubertazzo L[13], Vercellone A[14], Degli Esposti L[1]

[1]CliCon S.r.l. Health, Economics & Outcomes Research, Bologna, Italy, [2]ASL BAT, Trani, Italy, [3]Azienda ULSS8 Berica, Vicenza, Italy, [4]ASL Vercelli, Vercelli, Italy, [5]USL Umbria 2, Terni, Italy, [6]ASL Viterbo, Viterbo, Italy, [7]ASL Foggia, Foggia, Italy, [8]Azienda ospedaliero-universitaria Mater Domini, Catanzaro, Italy, [9]ASL Roma 6,Albano Laziale, Italy, [10]ASL Frosinone, Frosinone, Italy, [11]Direzione Generale per la Salute Regione Molise, Campobasso, Italy, [12]ASL 3 Azienda sociosanitaria ligure 3, Genova, Italy, [13]ASL Roma 4, Civitavecchia (RM), Italy, [14] ASL Napoli 3 SUD, Torre del Greco, Italy

CONCLUSIONS

This study showed that the main factors for predicting the prognosis in patients with COVID-19 are the **development of pneumonia and age**. The combination of the other features further helps the model identify critically severe patients by analyzing their recent clinical features. The obtained model could predict the progression of the disease in high-risk patients by analyzing the information present in the administrative flows. The model will be further improved through a feature selection process to increase accuracy and identify other important factors potentially predicting the severity of SARS-CoV-2 infection.