A machine learning approach for identification of gastrointestinal predictors for the risk of COVID-19 related hospitalization

General comments:

The manuscript is aimed to explain the SARS-Cov-2 predictive factors considering gastrointestinal and liver problems as symptoms for the hospitalizations. The paper is written in a clear and correct language. However, there are fundamental problems as explained below.

1) It is not wise to factor out a dominant predictor. Clustering the factors, for example, considering two or more factors together in different combinations may provide better results than factoring out a single dominant predictor. Plus, the difference of the ROC obtained is, as outlined in lines 188-190, only 3%, which is very small to arrive at such bold conclusion.

2) The difference in patients admitted to hospital and discharged home for AST is 0.74 – 0.56 = 0.18, for ALT, it is 0.61-0.48 = 0.13. This implies the mean values are almost equal showing that AST is not conclusive enough to factor it as an important predictor than ALT.

2) Most of these symptoms are also dependent on other environmental and epidemiological factors unique to a specific type of patient. So, there is no preferred predictor for this specific type of disease.

3) Some of the points that are inconclusive based on available data but mentioned in the paper. For example, on Line 161, and Table 1 the total number of female and male patients considered is 347 as opposed to 352 as explained in the "detailed comments" section below.

Considering, the above concerns and the following detailed comments, the paper needs to carefully address these issues before it is considered for publication.

Detailed comments:

Line 96: "2 distinct …" may be re-written as "Two distinct…". Starting sentences with numbers may create confusion since the authors also used numbers for citations.

Table 1: In the "Gender" row and "Outpatient test center: SARS-CoV-2 negative" category, the number of female, 206 and male, 141 add up to 347 but the number, n of the total negative outpatient number is 352. What is the reason for this discrepancy?

Line 161: “There were no significant differences based on sex.” The data isn’t shown for the gender category. So, how do we know this?

Lines 166-168: “Comparing SARS-Cov-2 negative and SARS-CoV-2 positive participants the presence of these symptoms has been more than three times higher in the positive group than in the negative one.” This sentence doesn’t seem to be supported by data. According to Table 1: The number of positive group, who are presented with the symptoms are 122 and the negative 101. They are almost equal and the difference is NOT a multiple of three.

Lines 188-190: The AUC values for the two curves is 0.799 and 0.76, the difference of which is 0.033 = 3%. So, can these be conclusive enough to quantify the dominance of the liver enzymes? Also as explained here,

Lines 59 – 61: “Furthermore, using machine learning random forest algorithm, we have identified elevated AST as the most important predictor for COVID-19 related hospitalizations.”