Question 1

The aim of this essay is to evaluate Brinati et al's attempt to diagnose COVID19 using machine learning.

By the time Brinati's paper was received (April 2020) there were 2 million confirmed COVID19 cases and almost 200,000 deaths (1). These numbers have increased significantly since then, as there are currently 148 million confirmed cases and 3 million reported deaths due to the pandemic. (2) PCR was considered the golden standard test for confirmation of the infection, but the paper argues that a better alternative is needed. The downsides of PCR are false-negative rates of 15-20%, long waiting times (3-4 hours), shortage of tests and the need for laboratories and trained personnel. (1) Because of these reasons, the authors attempted to create a test that generates output instantly, has better false-negative rates and does not require medical equipment.

The dataset was made available on Zenodo and is shown in *Figure 1*. (3) It consists of 279 patients admitted to the hospital between February and March 2020. The observations include the patient's age, gender, values from routine blood tests and the result of PCR. The dataset is imbalanced towards positive classes as 63% of the observations have a positive PCR result. First of all, it is important to note that the data sample is too small for feeding it into prediction models as 300 people definitely cannot represent the general population. It can be seen from *Figure 2* that the mean of the variable age is 61.3 and the median is 64, which indicates that the dataset consists of data of senior people. Moreover, the violin plot in *Figure 3* indicates that the data is more concentrated towards people older than 50. A general sample needs to be distributed equally across ages as variables derived from blood tests should be different in younger people. Another important comment that can be made about the data is that the authors used PCR results as a dependent variable in machine learning models, which means that it was assumed that PCR is always true, even though it has a false negative rate of 15-20% (sensitivity is 80-85%). The dependent variable is supposed to represent a general truth not bias. A way to overcome this might be having more than one PCR test done for each observation in order to reduce the error. Another study tried to overcome this issue by using both PCR and CT scan as a dependent variable in COVID19 diagnosis using machine learning. (4) Moreover, viruses are mutating in time so new strains are born such as the more contagious UK variant. (5) This variant might have different patterns in blood tests for predicting the positive or negative class, so as long the data is not constantly updated and fed with observations of the new strains of the virus, the machine learning model should not be used as an alternative method for COVID19 diagnosis. An improvement suggestion for the data used in this study could be using symptoms as a categorical predictor, as Mei and Lee did with their multi-layer perceptron model. (6)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	GENDER	AGE	WBC	Platelets	Neutrophils	Lymphocytes	Monocytes	Eosinophils	Basophils	CRP	AST	ALT	ALP	GGT	LDH	SWAB
2	M	56	2,9	128	1,9	0,8	0,2	0	0	29	36	18	43	21	257	1
3	M	56	3,5	151	2,1	0,9	0,4	0	0	16,5	25	14	50	17	207	1
4	M	72	4,6	206						193,7	31	22				1
5	M	72	16,5	316	14	1,2	0,3	0	0	318,7	96	33	80	42	651	1
6	M	77	4,9	198												1
7	M	77	3	162	2	0,4	0,4	0,1			21	10			220	1
8	M	74	5,3	189	3,3	1,3	0,7	0	0	4,6	16	13		27	117	1
9	M	74	5,2	144	4,2	0,6	0,4	0	0	104	91	131			391	1
10	M	75	11,6	123	10	0,8	0,7	0	0	244,6	53	37	43	27	439	1

Figure 1 (3)

Table 2 Descriptive statistics for the features considered in the present study

Feature	Mean	Std	Median	Kurtosis	Skewness
Age	61.3	18.5	64	-0.1	-0.5
Leukocytes (WBC)	8.5	4.8	7.2	2.3	1.5
Platelets	226.5	100.8	205	1.8	1.1
C-reactive Protein (CRP)	91.1	93.5	57.2	1.9	1.4
Transaminases (AST)	54.2	57.4	37	28.8	4.6
Transaminases (ALT)	46.6	47.1	33	11.8	3.1
Gamma Glutamil Transferasi (GGT)	82	128.8	48	14.8	2.6
Lactate dehydrogenase (LDH)	378	212.9	328	12.6	0.7
Neutrophils	6.6	4.36	5.3	3	1.6
Lymphocytes	1.2	0.7	1.1	16.1	2.7
Monocytes	0.6	0.4	0.5	8.2	2
Eosinophils	0.05	0.1	0	48.7	5.5
Basophils	0.01	0.03	0	14.2	3

Figure 2 (1)

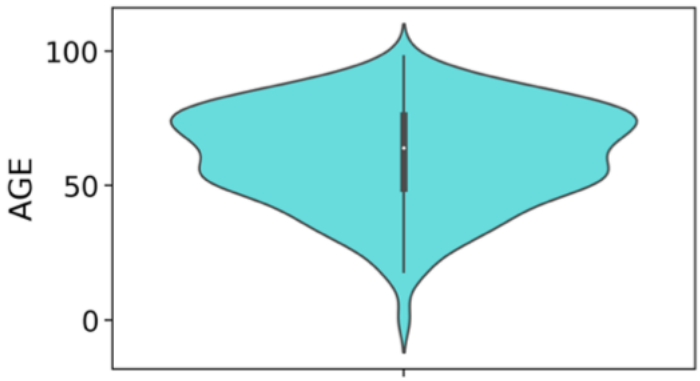

Figure 3 (1)

In the study, data manipulation was performed by handling missing values using multivariate imputation by chained equation (MICE), nested cross-validation and removing gender from the predictors as it was found of negligible predictive value. In the MICE process, each feature with missing values is modeled as a function of the other features. (7) As the data only consists of 279 observations, there is the danger of overfitting – creating a hypothesis that does not generalise enough and is not accurate on new examples. Russell and Norvig explain that a way to address this is using cross-validation, by using different subsets of the data when splitting into training and testing. (8) The inner loop in the nested cross validation allowed the finding of the optimal hyperparameters (for example, number of estimators or criteria for random forest)In inductive learning, the simplest hypothesis consistent with the data is preferred, according to Ockham's razor. (9) Therefore, a possible improvement in the

stage of data manipulation is eliminating highly correlated variables from the set of predictors. In *Figure 4* it can be observed that some variables are highly correlated: Transaminases (ALT) with Transaminases (AST) or with Gamma Glutamil Transferasi (GGT). In this case, two of these three attributes could be eliminated.

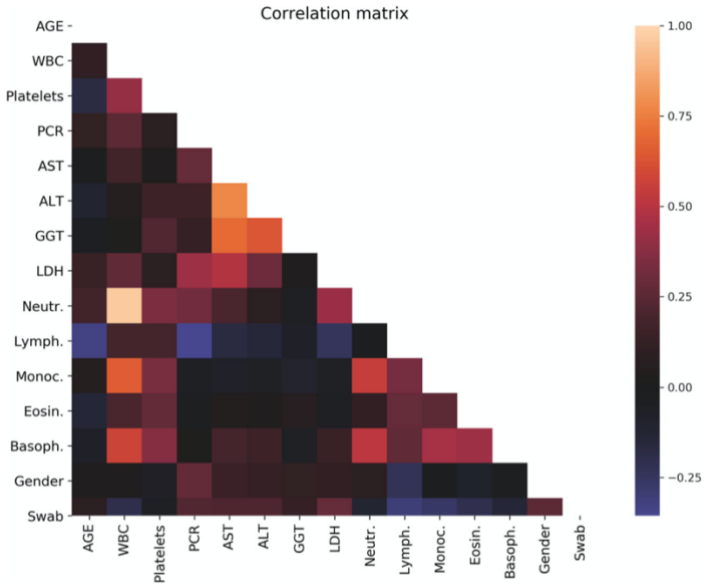

Figure 4 (1)

Seven prediction models were trained in the paper. The main quality metrics for assessing the models were accuracy and sensitivity (percentage of true positive out of all positives), as advised by the clinicians involved in the study. False negatives – infected patients allowed to go home – are considered more harmful than false positives. The authors selected 2 of the models they considered – random forest, which performed the best, and a decision tree, because the previous one could not be used as an interpretable model. (1)

Let us define these 2 models first. To start with, decision trees are a type of inductive learning, which means learning from data and creating a hypotheses h that agrees with the training set. In *Figure 5*, the 3 colored lines are possible hypothesis and the dots represent the training set. Decision trees can also be classified as supervised learning – the training examples have both an input and an output as opposed to unsupervised learning where no explicit feedback is given. In a decision tree, each leaf represents an output value to be returned, and each internal node verifies the value of an input attribute, while the branches represent the possible values. (9)The algorithm for training a decision tree is built recursively in a top-down manner: it goes back to the previous level of recursion if there is no more input data, or it returns a leaf node if the values in the training data belong to the same class, or it returns the class with the highest plurality value if there are no more attributes to be tested or it uses the information gain as a metric for selecting on which attribute to split. (10) The information gain is calculated by subtracting the entropy after splitting on an attribute from the entropy before splitting. If we consider a binary classification problem with p representing positive values and n representing negative values, the entropy of a node is Entropy(p/(p+n), n/(p+n)) = - p/(p+n) log (p/(p+n)) - n/(p+n) log (n/(p+n)). Each node

has the weight of the branch which is how many examples it selected out of the total number of examples before the split.

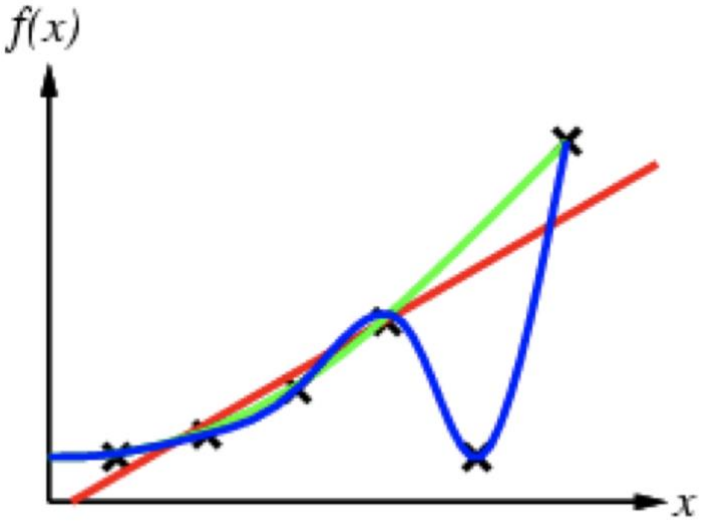

Figure 5 (9)

After having defined decision trees, it is easier to understand random forests, as the algorithm creates more trees that will vote on the outcome. The output is going to be a probability score for both classes. This score is going to be assigned to one of the classes by comparing it to 2 thresholds, $\alpha, \beta \in [0,1]$ and to the class that has a higher probability. M random features and N random attributes are selected for each tree. In the paper, the split is done using Gini impurity. It represents the probability of a randomly chosen element to be incorrectly labeled if it was randomly labeled according to the distribution label in the set. This probability is given by the formula $I_G(p) = \sum p_i (1-p_i)$. The split that minimizes the Gini impurity is selected. (10) The number of estimators is specified – it means the maximum number of trees in the forest could be 100. The feature importance is defined as the total normalized reduction across the decision trees in the random forest to the variance of the target feature. The most important features found were AST and Lymphocytes. The authors of the papers could have considered eliminating the last 5 features, as their score is below 0.05. This could have improved the model performance and satisfy Ockham's razor (9), the principle of finding the least complicated hypotheses.

A modification of the random forest model was also considered: a three way random forest (TWRF) classifier allows the model to abstain on instances for which it expresses low confidence – and thus achieves higher accuracy but lower coverage. The accuracy and sensitivity of the traditional random forest model are 82% and 92% while the accuracy and sensitivity of the TWRF are 86% and 95%. However, these models achieve this sensitivity on a test set of 55 observations (20% of the full data). PCR has sensitivity of 80-85% but PCR is the golden standard for COVID19 testing, which means it was tested on a very large and general sample of the population. This implies that a sensitivity of 92% is not necessarily more accurate than the PCR.

As mentioned before, a decision tree was used as an explainable model because the random forest is not able to provide interpretable insight into how predictions are made, as it is an ensemble model averaging the predictions from more trees. The authors claim that the Decision Tree model is used to approximate the decision-making steps of the Random Forest and that even though the performance of the model is lower than the Random Forest's, the Decision Tree can be used as a decision aid by clinicians. A very important point to make here is that it is quite inappropriate to say that a Decision Tree built on the full dataset can approximate the decision-making steps of a random forest of 100 decision trees from different subsets of the dataset. The splits are going to be different in the random forest because subsets of the data are taken. While some of the variables' importance coincide in the 2 models, as it can be observed from Figure 7 in the paper (feature importance in random forest) and Figure 8 in the paper (decision tree output), it is not the case for all of them. LDH, which is the third most important variable in the random forest, is not even a splitting criterion in the decision tree. What is more, it is not such a good idea to use the Decision Tree as a decision aid by clinicians as Figure 5 in the paper, the ROC curve, suggests that the Decision Tree performs the poorest out of the 7 trained models. The ROC curve shows the relationship between true positive and false positive values in a model.

To sum up, the paper could benefit from some improvements, including adding more and more generalised data from people of various ages and from more COVID19 tests as predictors, eliminating highly correlated variables and model features that have the least importance in order to satisfy Ockham's razor, and in the case of using a Decision Tree as explainable AI, improving its performance first and avoid saying it can estimate the decision-making process of a random forest.

Word count: 1792 words

References

1. Brinati D, Campagner A, Ferarri D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study [Internet]. 2020. Available from: https://moodle.ucl.ac.uk/pluginfile.php/3894560/mod_resource/content/8/Brinati2020_m2.pdf

2. COVID Live Update: 148,480,035 Cases and 3,133,637 Deaths from the Coronavirus - Worldometer [Internet]. Worldometers.info. 2021 [cited 25 April 2021]. Available from: https://www.worldometers.info/coronavirus/?utm_campaign%3DhomeAdvegas1

3. [Internet]. Zenodo. 2020 [cited 24 April 2021]. Available from: https://zenodo.org/record/3886927#.YIcMyS2w1R0

4. Li W, Ma J, Shende N, Castaneda G, Chakladar J, Tsai J et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2020;20(1).

5. What are the Indian, Brazil, South Africa and UK variants? [Internet]. BBC News. 2020 [cited 22 April 2021]. Available from: https://www.bbc.com/news/health-55659820

6. Mei X, Lee H. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature Medicine. 2020;.

7. Kas K. COMP0014 Lecture 5: Learning. Lecture presented at; 2021; University College London

8. Russell S, Norvig P. Artificial Intelligence: A modern approach. 2010.

9. Joh D. COMP0014 Lecture 5: Learning. Lecture presented at; 2021; University College London

10. Cha Y. COMP0014 Practical 7: Learning. Lecture presented at; 2021; University College London

Question 2

a) As discussed by Branchman (2002), reactive, deliberative and reflective are three types of processes that intelligent systems base their interactions with the environment on, by using sensors or acquired knowledge, and which can happen in parallel, alongside each other. They differentiate by the degree of awareness of the system and the level of input data abstraction involved in the process and are ranked from the lowest to highest in the order above. A reactive process, which is also described as “autopilot”, can be either an innate reflex or learned, but relies solely on “raw” sensor data, without reasoning about it – doing without thinking. Deliberative processes are the ones where systems also use knowledge in addition to sensor data in order to reason about a matter, and as the name suggests, consciously to a certain degree, with the intention to achieve its goal. It is what goes by the name of “thinking”. In a reflective process, the highest level of abstraction, the system would be aware of its own “thinking” processes and will try to reason about them. For instance, it is given the case of solving a problem and stopping the cognitive process to look for other approaches.

   An example of reactive behaviour in a more evolved version of Samsung’s chef assistant robot might be judging whether meat is already well done based on some kind of chemical “sense of taste”, even if the recipe states it needs to cook for longer, in order to achieve better precision.

   In its current state, the robot is able to download recipes from the internet and execute them. More deliberative behaviour could involve improving these recipes by “thinking” – for instance, it could alter steps and ingredients to adapt to different dietary requirements, such as removing allergens.

As Samsung's bot is meant as an assistant to the human chef, in a more advanced version it could implement a reflective behaviour that would increase the efficiency of chef-assistant cooperation. The bot should be aware of its limitations, two robotic arms fixed on a wall, and reflect on how can make the process of cooking more efficient. To be more specific, it can take either feedback from the chef regarding the quality of the final product, or compute the time required for preparation by itself and then, comparing the feedback tasks completed by it with feedback from the tasks completed by the chef, it could learn how to split the tasks better next time. For instance, it could learn that the chef is better at picking ingredients, while its robotic arms can cut faster and more precisely, or can mix smoother.

b) Symbolic AI means reasoning with a set of rules, as in an IF-statement, an induction process that reaches a conclusion by following some premises or even a deduction process that looks for evidence that a statement is true. On the other hand, subsymbolic AI is not encoded as a set of rules, but as a set of numerical patterns that use induction to reach an output. The most popular example for this is a neural network, as in this black-box model, the input is learned using neurons and connections between neurons to reach a conclusion. Symbolic Ai is preferred when a problem can be solved by human-mind and also the thought process can be explained so that the decision can be explicitly backtracked. Subsymbolic AI is preferred when a verbal justification is not required and when the human mind can solve a problem but it is very difficult to explain how, for example when recognising faces of familiar people. One application best suited for symbolic AI could be deciding whether to give loans, as a person refused should be able to request a reason. A symbolic AI such as a decision tree is able to explain the "thinking process". This is considered a great strength. At the same time, subsymbolic AI is disadvantaged in this case, as it is impossible to backtrack its decision process. One application best suited for subsymbolic AI is object recognition. This kind of AI has the advantage that it is trained using a set of images to distinguish between a cat and a dog for example. The subsymbolic AI aim is to develop an internal representation of the essential features of both classes. However, object recognition is considered to be a weakness for symbolic AI. Trying to find a rule to recognise a chair can be challenging: a chair definition can be that it is a portable object, has a horizontal surface suitable for sitting and has a vertical surface positioned for leaning against. If the system respects this definition it will fail to correctly label armchairs as they do not have any legs and are not portable. This is because it is extremely difficult to build a definition that incorporates all the features of an object.