Dear Editor Dr. Sabine Loos,

We would like to thank you and Reviewer 1 for re-evaluating our revised manuscript. We appreciate the time and effort that you have dedicated to providing feedback. We hope that our revisions will satisfy the proposed requirements for successful publication. Please see below for a point-by-point response to your and Reviewer 1's comments and concerns.

Yours sincerely,
Oya Kalaycioglu, PhD, on behalf of all authors

EDITOR'S COMMENTS

1. Explaining methods for quantifying variable importance – Because the motivation of this paper is to understand which factors influence social vulnerability by using Machine Learning models, it is important to more thoroughly understand how variable importance is quantified for all models used beyond the references included in the text. Please add more description in the Methods section.

Authors' response: Thank you for this valuable suggestion which helped us to improve the methodology section of our manuscript. We have included a new sub-section "*3.8 Variable importance analysis*" which provides details about how the ML algorithms and logistic regression determine the importance of variables.

2. Earthquake-specific versus all-hazards vulnerability – Additional justification is required that this model is relevant for all-hazards rather than specific to earthquakes, especially since the dataset that was used is an earthquake-specific household survey. Corresponding revisions are necessary in the manuscript.

Authors' response: Thank you for this comment. In the revised version, we have made the relevant revisions to make clear that our model is applicable for hazard related vulnerability. Cutter et al. (2003, Social Vulnerability to Climate Variability Hazards: A Review of the Literature, p.23) reported that *"… the accepted theoretical understanding that social vulnerability is independent of hazard type. Zones of differential exposure to any or all hazards combine with SoVI to create place vulnerability (for example see Borden et al. 2007; Burton and Cutter 2008; Wood et al. 2009)".* Please see relevant information in subsection *3.2.2 Construction of SoVI*, p.9. L239-242. Therefore, we assumed that earthquake related data collected in household survey and the indicators used for SoVI in Istanbul are applicable to other hazard events as well (please see L214-217 and L268-270).

3. Justification of ML models' performance – Can the authors justify the reported model performance for the ANN (and other ML models used) and why this is acceptable, perhaps by comparing to similar studies?

Authors' response: Following your and the Reviewer 1's comments on our model accuracy, we have revised the sub-section "*5.1 The selection of the optimal ML method*". In the revised

version, we have now included the discussion on our optimal ANN model's (and other models') performance. We compared the performance metric values we obtained with our ML models to the acceptable values in the literature. We considered our proposed ANN model to have a good discriminative ability (AUC>0.80) according to Hosmer et al.'s (2013) criteria. We also considered the accuracy of our optimal ANN model (which is 73%) to be acceptable as the value is halfway between 50%, which is useless, and 100%, which is perfect (Power et al., 2013). There is a limited number of studies in the literature that have used ML to predict hazard-related social vulnerability and reported performance metrics. We also included the performance metrics they have reported. We discussed that although our models' accuracies were relatively lower compared to other studies which assessed social vulnerability to hazards with machine learning techniques, our approach can be useful for decision-makers to take immediate action for the most vulnerable households. We also noted that, our models can further benefit by incorporating more predictor variables.

REVIEWER 1' COMMENTS

Medium and Minor Issues

1. L23: "CART" is short for "classification and regression tree". Please use either "classification tree (CT)" or "classification and regression tree (CART)" to avoid confusion. The same for the rest of the manuscript.

Authors' response: We revised the terminology and used "classification and regression tree (CART)" throughout the text and made the correction in L23.

2. L40: When used as a countable noun, "vulnerability" usually means a weak link, a loophole, a fragile element, etc., of a system that may be exploited by hazardous agents to result in loss to the system. Is that what the authors mean here? If not, please use the uncountable version of the noun "vulnerability".

Authors' response: We corrected the typo in the revised manuscript and used uncountable version of the noun vulnerability in L40.

3. L86: I suggest removing "in fact" because the statement here is still about an intellectual guess or belief.

Authors' response: We agree with the reviewer's suggestion and removed "in fact" in L86.

4. L150-151: Very large earthquakes (over Mw7.0) do seem to be rare for Istanbul. However, earthquakes with a magnitude 4 or above can still cause significant damage to communities (see, e.g., Wang and Sebastian 2022 https://doi.org/10.5194/nhess-22-4103-2022). These

earthquakes shouldn't be rare at all according to the estimated 100-year return period for an earthquake with Mw7.0 and above around Istanbul. In addition, as the manuscript has changed its focus from earthquake to all hazards, the large hazardous events should be more frequent than merely large earthquakes. Even the authors themselves mention on L203-204 that the study area "is in a region that is prone to natural hazards where a large-scale disaster happens every seven to eight years (Baris, 2009)". Moreover, at the household level, many families do not have to wait for a large-scale disaster to occur before experiencing loss unfortunately. More impacts to households are likely to be caused by the much more frequent smaller-scale hazardous events that may not even be considered or defined as disasters.

Authors' response: We thank the reviewer for this comment. We agree with the reviewer that smaller-scale hazardous events can cause losses at the household level as well. As this paragraph relates the pros and cons of using historical data to calculate SoVI, we mentioned that when catastrophic hazard occurrence is rare, the policy-makers can underestimate the impacts of a major hazard event, if they rely on historical data from the smaller-scale hazardous events where the losses are much less due to infrastructural investments. For example, in İstanbul – Türkiye, primarily an earthquake-prone zone, using empirical loss data of frequently occurring small-scale hazard events may mask the possible impacts of a major earthquake (over 7.0 Mw), which is rare due to historical records. Please see the relevant revision in L148-155.

5. L202-210: This paragraph is inappropriate for hazards in general. Please revise it.

Authors' response: We thank the reviewer for this suggestion. We revised sub-section *3.1 Study Area* to reflect hazards in general and provided more relevant information about Istanbul (L194-211).

6. L213-214: Since the focus is on hazards in general instead of earthquake now, I suggest the authors explain a little bit regarding why the survey conducted by an earthquake-related organization can be used for all hazards.

Authors' response: We made clear that the household survey data that we have relied on is collected by İstanbul Metropolitan Municipality (IMM) in 2017 to assess *disaster-related* social vulnerability of the households in İstanbul. The variables used in this research were in line with the social science and disaster literature, where such research is focused generally on the social factors that increase or decrease the impact of specific hazard events on the local population (L214-217).

7. L231-234: There are grammatical problems associated with this sentence "It considers … and socio-economic status". Please modify it.

Authors' response: To modify this sentence, we advised to a native English speaker. Following his suggestion, we modified this sentence. (L233-235).

8. L303-305: The authors claim that for "different tuning parameter alternatives, the choice of the optimal tuning parameter was determined by the largest area under the curve (AUC) value of the receiver operating characteristic (ROC) curve using the automated grid search". However, in the supplementary file 3 p. 2, the authors clearly state that the parameter K of KNN is "determined with the square root of the number of points in the training data set". These two statements are inconsistent with each other. Why? Moreover, the parameter **ntree** of RF is also determined arbitrarily by the authors without grid search.

Authors' response: In the previous version of the supplementary material, we made a typo regarding the explanation for determining the tuning parameters. As we have stated in the manuscript, we used automated grid search to find the optimal tuning parameters. Accordingly, we revised the supplementary file 3. p.1-2, and also clearly stating that we determined tuning parameters with grid search for all methods in the manuscript (L291-292).

9. L403-404: The sentence "The prevalence … among 41,093 households" needs to be modified for all hazards.

Authors' response: We revised the sentence accordingly (L413).

10. L422-424: Sensitivity and recall are the same thing. Positive prediction value and precision are the same thing.

Authors' response: We have removed the terms which imply the same metric in the manuscript (L434). However, we kept both terminologies in the R-shiny app, to help the readers who are familiar with different terminology.

11. L435: Fig. 3 may not be friendly enough to colorblind readers.

Authors' response: We thank the author for this suggestion. We have revised the colors in Figure 3, according to R color guidelines which provide a color palette for colorblind readers.

12. L468-472: It is still unclear in the manuscript how the relative importance of predictors is measured. What methods or algorithms do the authors use for quantifying predictor importance?

Authors' response: We thank the reviewer for this suggestion, which helped us to improve the methodology section of our manuscript. We have included a new sub-section "*3.8 Variable importance analysis*" (p15, L370-390), which provides details about how the ML algorithms and logistic regression determine the importance of variables.

13. L475: In Fig. 4, it is unclear whether it is the size of the circle or the hue that is supposed to correspond to the number next to the circle in the legend. Also, the scale of the circle size does not cover the small value as for debt in Fig. 4A. In addition, what do the colors of the bars in

Fig. 4b mean? If they indicate the variable importance in terms of percentages, then these colors provide redundant information that is confusing.

Authors' response : We thank the review for this suggestion. We corrected the circle size of the "debt" variable. Also, we want to clarify that both the size and the color of the circles in Fig. 4A represent number next to the figure legend. We also plotted this figure using either only varying the size or color of the circles. But as the circles were overlapping for some variables, it was hard to interpret the difference between the variables when only one of these strategies (i.e., size or color) was used. Therefore, for Fig. 4A we kept the version where both size and colors varied. However, for Fig 4B, we agree with the reviewer and in the revised version we used the same color for the bars. In the previous version, the gradient color was used as an indication of lower percentage, but as these percentages were already provided in the figure we agree that varying colors were confusing.

14. L506-507: I find it difficult to see the connections between this sentence "For many decades…derived variables (Di Franco and Santurro, 2020)" and the rest part of this paragraph. I suggest the authors make modifications to the paragraph.

Authors' response: To focus on discussing our findings from our ML models, we removed the first sentence starts with a comment on data analysis and social sciences. We revised this paragraph (L514-523)

15. L513-521: This paragraph reads awkward. What the main point is here is unclear. The authors need to revise the paragraph. Also, it is dangerous to assume that the trained non-linear structure of the neurons of an ANN represents well the relationships between the input variables. Every training may result in a totally different internal structure of the ANN.

Authors' response: Following the Editor's and Reviewer 1's comments, we have revised the sub-section "*5.1 The selection of the optimal ML method*". In the revised version, we have now included the discussion on our optimal ANN model's (and other models') performance. We replaced this paragraph with a new paragraph comparing our results from the ANN model with the results of different studies. For these comparisons we focused on AUC and accuracy as an indication of the model performances (L524-539).

16. L530-531: Why is it important whether the training data has to be balanced? If the identification of high social vulnerability is preferred, why don't the authors use a reversely imbalanced training data to boost the sensitivity, etc., even more?

Authors' response: The consequence of ignoring the issue of imbalanced data is to over-predict the class with higher frequency (Esposito et al., 2021, doi.org/10.1021/acs.jcim.1c00160 J). This increases specificity, and therefore reduces sensitivity. In our case, that results in over-predicting low vulnerability group, thus increasing specificity. However, our aim was to increase sensitivity to identify the households with high social vulnerability more accurately. The sensitivity increased when we used under sampling, which discards data points from the

majority class (i.e. low vulnerability group) at random until a more balanced distribution is reached while training models. Please see L546-552 for relevant explanation.

17. L538-549: What is the main theme of this paragraph? Are the authors trying to discuss the methods for measuring importance of input variables or the important input variables identified in their study? Also, as I have asked previously, what is actually the method or algorithm that the authors use for their quantification of variable importance?

Authors' response: We revised this paragraph. We know included a sub-section "*3.8 Variable importance analysis*" which provides details about how the ML algorithms and logistic regression determine the importance of variables. In this section, we know only discuss the important input variables identified with our models.

18. L538-585: The writing of these 3 paragraphs needs to be improved.

Authors' response: We have revised these 3 paragraphs. (L559-598).

19. L566-568: It does not make sense that prevalence of social security in low vulnerability households is lower than in high vulnerability households.

Authors' response: It was a typo and we corrected in the revised manuscript. (L581-582)

20. L569-585: The material in this paragraph involving earthquake needs to be modified to fit the tone for all hazards.

Authors' response: We have edited the paragraph to be applicable to all hazards. (L583-598)

21. L624: What are "hazard risks"? I suggest the authors use "hazards" here instead.

Authors' response: It was a typo and we corrected in the revised manuscript. (L633)

22. L641-642: Why "fault lines"? Is the manuscript supposed to be for all hazards now?

Authors' response: We removed the sentence, which emphasises the fault lines, as the manuscript is focused on all hazards now.

23. The authors' final ANN model has an accuracy of less than 75%. That is quite low for identifying high vulnerability households. Isn't this an obvious limitation? If I were the public official to look at vulnerable households, I would definitely need a prediction model with accuracy over 90%, or even 95% or 99%. How could I afford to miss those actually vulnerable households while providing a lot of resources to households that are actually not highly vulnerable?

Authors' response: Our final proposed ANN model had an AUC >0.80 which indicates a good level of discriminative ability between households with high and low social vulnerability according to Hosmer et al.'s (2013) criteria. We also considered the accuracy of our optimal ANN model (which is 73%) to be acceptable as the value is halfway between 50%, which is useless, and 100%, which is perfect (Power et al., 2013). However, the accuracy of our models were relatively smaller compared to the similar studies which use ML models to predict SV (Abarca-Alvarez et al., 2019, doi.org/10.3390/ijgi8120575; Alizadeh et al., 2019, doi.org/10.3390/su10103376). That was due to the fact that, we used quantifiable household data as our aim in this manuscript was to present an optimal modelling strategy capable of processing readily available large databases. Our approach can be useful for decision-makers to take immediate action for the most vulnerable households, and further can be enhanced by incorporating more predictor variables. We have revised the sub-section "*5.1 The selection of the optimal ML method*" accordingly and added the limitation of relatively low model accuracy to section *"6* limitations and recommendations" in the revised manuscript. (L654-656).

24. Supplementary file 3, p.2, 4. SVM: Why do the authors use the radial basis function (RBF) kernel? A linear kernel can correspond better to a hyperplane in a vector space with the same number of dimensions as the number of input variables. When an RBF kernel is applied, it is equivalent to transforming the original vector space into a vector space with an infinite number of dimensions. Such a transformation can be demonstrated with the application of a Taylor expansion when we are using the Lagrange multipliers to solve the optimization problem for calibrating the SVM.

Authors' response: For training data with SVM, we have fitted three possible versions of kernel functions which are radial kernel, polynomial kernel and linear kernel. As mentioned in supplementary file 3, p.2, 4. SVM, line 6-7, radial kernel is used in our study as it provided larger AUC compared to using linear or polynomial kernel functions. Thus, as we obtained the highest discriminative ability between the households with high and low SV when we used RBF function.

25. Supplementary file 3, p.2, 7. ANN: ANN is not necessarily non-linear. When all the activation functions are linear, an ANN is equivalent to a linear model. ANN is also not necessarily based on deep learning. Deep learning involves an ANN with at least 2 hidden layers. If the authors only use 1 hidden layer, that is not deep learning. Also, what activation functions do the authors use for their multilayer perceptron ANN model?

Authors' response: We thank the reviewer for this comment. We made the following revisions in the supplementary file 3, p.2, 7. ANN method:

> **7. Artificial Neural Network (ANN),** which is capable of learning any non-linear function, is ~~a powerful non-linear learning technique~~ created by imitating the functioning of the human brain and transferring it to the computer environment, which is first proposed by McCulloch and Pitts (1946). ~~It is based on deep learning mechanism that works with iterative propagation algorithms.~~ The mechanism operates with

artificial neurons that form input and output neurons and a hidden layer(s), which is frequently used for the data set that cannot be separated linearly. When there are multiple layers between the input and output layers, ANN is as a deep neural network (DNN) (Schmidhuber, 2015). Although computationally expensive, it is successful to detect complex nonlinear relationships between variables. We used sigmoid/logistic function as the activation function, which is commonly used to add non-linearity to an ANN model. Additionally, ~~We used~~ the multilayer perceptron choice was employed for ANN model by adapting **nnet** method in **caret**, which contains two tuning parameters: number of neurons in a hidden layer and decay parameter that controls initial weights for input neurons.