Response to Reviewer 1 Comments

In this manuscript, the author aims to predict cardiovascular risk groups using machine learning by extracting data that can be measured with a smartwatch from the Korean National Health and Nutrition Examination Survey. Although this research is beneficial to the Korean society, there are some issues regarding the machine learning model building method.

Comments

1) This title misleads readers into thinking that the prediction model was built based on data collected from smartwatches. Please change the title appropriately.

➔ Following the reviewer's comments, I have modified the title using 'the Korea National Health and Nutrition Examination Survey'.

2) Abstract

The authors should clearly state the sources of their information, the Korean National Health and Nutrition Examination Survey.

➔ As the reviewer commented, I have stated the sources of information, the Korea National Health and Nutrition Examination Survey to the abstract. Also, the Korea National Health and Nutrition Examination Survey was added to keywords.

3) 3.1. Data source

"The final 6,170 people were selected for analysis of this study after removing missing and abnormal values from the latest 2019 survey results." Please specify the number of subjects in "The Korea National Health and Nutrition Examination Survey", the criteria for selection or exclusion of subjects in this study, and the definition of abnormal values.

Please illustrate each exclusion step as a scheme with the number of remaining subjects.

➔ Following the reviewer's comments, I have explained the number of subjects in "The Korea National Health and Nutrition Examination Survey", the criteria for selection or exclusion of subjects in this study, and the definition of abnormal values in Section 3.1 (page 3).

4) 3.1.Data source

"The reason for linking the prevalence of diabetes and the presence of cardiovascular disease is the result of research showing that if blood sugar continues to be high due to diabetes, blood vessels are damaged and various cardiovascular diseases are accompanied [23]."
Please add a note regarding the integration of dyslipidemia.

➔ Following the reviewer's comments, I have added previous study on the relationship between dyslipidemia and cardiovascular disease in Section 3.2 (page 3).

5) 3.1.Data source
"As mentioned above, Samsung Electronics' smartwatch can measure heart rate, oxygen saturation, stress index, blood pressure, and electrocardiogram As mentioned above, Samsung Electronics' smartwatch can measure heart rate, oxygen saturation, stress index, blood pressure, and electrocardiogram, and blood glucose can be measured in the second half of 2021."
It is unclear what "stress index" means. Please mention the content of this variable.

➔ To address the reviewer's concerns, by dividing the sentence into two, the first sentence explained the aforementioned part, and the second sentence explained the already provided functions. Then I have added a description of the stress index in Section 3.2 (page 3-4).

6) Table 1.
In order to use BP1 (perceived stress) as an explanatory variable in machine learning, it is necessary to prove that it corresponds to the stress index displayed by smartwatch. Please insert an explanation on this point in the text.

➔ To address the reviewer's concerns, I have added that the subjective perceived stress level is used instead of the smartwatch's stress index because the perceived stress reflects the stress index displayed by smartwatch (page 4).

7) Please insert information about the details of each Variable shown in Table 1.
For example, please specify the number of men/women, the average and SD of blood pressure, etc.

➔ Following the reviewer's comments, I have added the details of each variable shown in Table 1 (page 5).

8) 3.3. Research procedure and 4.1.Performance evaluation index
In this manuscript, there is no description of the validation method for the prediction model constructed by machine learning. In machine learning methods, overfitting must be avoided by implementing some kind of validation method. Since the sample size in this study is large enough, it is recommended to separate the test

set group of about 20~30% from the training set group. After building the model with the training set, the prediction performance should be checked by the test set.

➔ To address the reviewer's concerns, In order to avoid the problem of overfitting the model, about 70 % (4,319 respondents) of the observations in the entire data set are trained as a training data set. And the performance of the optimal model is verified with the test data set of 30% (1,851 respondents), which is the remaining data set not used for model training. I have explained the validation method for the prediction model in Section 3.3 (page 5)

9) 4.3.1. Optimal model of support vector machine
Re-evaluate the model using the test set.

➔ Since support vector machine has the highest Accuracy, Recall, and F1 Score among the three models in the validation process for the prediction model in Section 4.2, I have re-evaluate the svm model using the test set in Section 4.3.

10) 4.3.2. features' importance
Describe the method for calculating the importance.

➔ Following the reviewer's comments, I have explained the method for calculating the importance in Section 4.3.2 (page 8). In addition, I have added the function name of R program.

11) Conclusion
"In this study, predictions were made for all data 341 without dividing the data sets into training and test, but since the prediction results differ In this study, predictions were made for all data 341 without dividing the data sets into training and test, but since the prediction results differ depending on the training and test data sets, further research is needed on the minimum number of training data that does not affect prediction performance."
Avoiding overfitting is essential in machine learning. A predictive model built without any validation is not a scientific achievement. Validation through the use of a test set is a minimum requirement for acceptance of this manuscript.

➔ To address the reviewer's concerns, In order to avoid the problem of overfitting the model, about 70 % (4,319 respondents) of the observations in the entire data set are trained as a training data set. And the performance of the optimal model is verified with the test data set of 30% (1,851 respondents), which is the remaining data set not used for model training. I have explained the validation method for the prediction model Then, since support vector machine has the highest Accuracy, Recall, and F1 Score among the three models in the validation process for the prediction model, I have re-evaluate the svm model using the test set.

I sincerely appreciate the reviewer's thoughtful comments that helped me improve the paper.