Peer review

Chintaluri, Anirudh, and Matthew Radzihovsky. 2025. “Machine Learning Models for Cardiovascular Disease: A Review on Early Diagnosis Performance.” *Journal of High School Science* 9 (1): 74–87.

I enjoyed reading the manuscript. It was well written and presented. However I have some points of concern that need to be addressed.

1.Please provide a confusion matrix for all the 5 classes of severity (predicted versus actual). The numbers in Table 4 do not enable deduction of how the accuracy, precision, recall…. are reflected across all the 5 severity categories. Also provide AUC (and if possible, PRAUC curves for the various ML models tested.

2.You state that “….However, detecting CVD itself can require the use of a multitude of equipment, tests, and analyses to properly diagnose one with CVD, with tests such as echocardiograms, stress tests, and coronary angiograms having median costs of up to $2588, $3230, and $9203 respectively in 2022……”. With any ML model, you would still need this data for the patient because the features from Table 1 still require cp, chol, restecg, thalach, exang, ca and thal. These features need for the patient to get an ECG, stress test and a cornonary angiogram (ca). Hence, the ML is still going to end up costing the patient the same money, time, clinic visits…. Please explain and discuss in the manuscript.

3.Please explain and present in Table 1 what the feature “thal” exactly means.

4.Provide the class distribution (showing the class imbalance) for the 303 instances divided into the 5 severity (0-4) categories.

5.What is the purpose of the last binary value column in Tables 2 and 3, when there is already an ordinal 0-4 severity of disease “True label” column. The binary label does not figure in the list of features in Table 1. Please explain in the manuscript.

6.You state that you normalized your data in a 0-1 range if it was quantitative. If that is so, then why are some chol, trestbps, age, thalach values negative after normalization?

7.Please report hyperparameters used for each lightweight model so that readers can replicate your work. Please provide a feature importance map or chart for each of the ML models.

8.Were the train-test splits stratified ? In other words, was every severity class evenly represented between the test and train categories ? If not, please explain what effect that would have on the models’ accuracy, recall… etc.

9.You also mention 5 trials of random test-train splits ? I am assuming this was for one ML model and represents something like a 5-fold cross-validation? Can you provide the maximum standard deviation for accuracy, recall and precision across the 5 trials for each ML model tested?

10.You state “……The recall and precision metrics used in this study were calculated through macro-averaging to take into account all five label categories……” Is this an arithmetic average? If not, can you either explain how this was calculated or provide an example?

11.You mention wearables. It would be interesting to see if you could use a correlation analysis or a PCA to only use features such as age, sex, trestbps, and restecg (which can all be accessed through a smartwatch or requires no doctor's visit) to still be able to predict CHD severity with sufficient accuracy. If possible, can you run such a 'wearable' ML model ?

12.I did not understand what the following sentence meant "…..Through hyperparameter tuning, we were able to determine which values are to be filled in depending on the KNN model with the highest accuracy with k=10 for ca and k=5 for thal….." What about the other features whose values were missing? Why do you need an ML model to fill in missing values ? I thought these were filled in using either the mean, median or mode. Please explain and discuss in greater detail in the manuscript.

13.You mention consistency and variability pertaining to the accuracy, recall, precision, F1 scores in Table 4 for each ML model. Why is consistency between these scores important ? I think you can delete Figure 2 because it gives you the same information as Table 4 and the differences are not enough to necessitate a column chart. Instead, if you want, you can add a standard deviation column in Table 4 to better illustrate your text results.

14.Where the reference lists more than 6 authors, the first 6 authors need to be presented followed by an et al. in the Reference section of the manuscript.

---

1. We have provided confusion matrices for each model as well as AUC values under the ROC curve (Figure 3). The confusion matrices shown are confusion matrices from each of the 5 runs from the 5 folds, added up together.

2. The claim about saving money specifically pertaining to diagnosis testing was misleading. The advantage with early detection with machine learning here is that it prevents more intense care in hospitals for patients, saving money from those procedures, not diagnosing the CVD itself.

3. Thal is short for Thalassemia, and the values in the dataset reflect the defect types. Included in Table 1.

4. Distributions for the class attribute before and after SMOTE are shown in Figure 1.

5. Originally we were going to look at how different binary labels were compared to multiclass, hence the "label" being replaced with "true-label" and the binary labels being "label." We decided not to do that and forgot to remove the binary label from the DataFrame, keeping all the attributes as they are meant to be. Corrected versions are displayed in Tables 2 and 3.

6. They are negative after normalization because for the attributes with discrete values, I did them from a 0-1 range, but for continuous values I used z-score. I did not explain this clearly enough in my manuscript. Included in Methods, part 2 (Preprocessing).

7. Included in the Methods section, part 3 (Model Development) in an ordered list.

8. No, the train-test splits were not stratified. Because the classes were perfectly balanced after SMOTE, there was no need for stratified splits. Because the size of the dataset is reasonably large (820), splitting it so that 20% of the data is for testing means that the effect of class imbalance should

be insignificant (Methods part 2, Preprocessing)

9. Yes, it is 5-fold cross-validation. It appeared that I have not done cross-validation properly in my initial submission, using five randomized train test splits instead of five folds from the dataset. Had to change up the procedure to make sure that this was correct, also changed the results and interpretability. Maximum standard deviation included in Results section.

10. Yes, macro-averaging is an arithmetic average, averaging the precision and recall scores based on the number of unique labels.

11. Included in Discussion section. I ran a Pearson correlation analysis and also ran a Random Forest model, informed the reader about how well it performs with the attributes not requiring specialized testing.

12. Explained in Methods part 2: Preprocessing. KNN imputation is another method of replacing missing values based on which value aligns best with the instances with missing data.

13. I did not get the consistency phenomenon after re-running cross-validation as previously mentioned in comment 9. I added standard deviation to each cell in Table 4 using a plus or minus, and removed Figure 2 from the previous submission.

14. Check references section, citation (5). Replaced all authors with et al.

Thank you for addressing my comments. I am happy with all of them except for 8.

1.For point 8 you state that, “No, the train-test splits were not stratified. Because the classes were perfectly balanced after SMOTE, there was no need for stratified splits. Because the size of the dataset is reasonably large (820), splitting it so that 20% of the data is for testing means that the effect of class imbalance should be insignificant (Methods part 2, Preprocessing)”. Howeever, does this not mean that there could have been an uneven distribution of class 1, 2, 3 and 4 severity distributed between Training and testing data sets? Please explain and discuss in the manuscript.

2.Although you have described pearson’s coefficients, did you perform any feature engineering as regards feature importance specifically for the different models used. For example, GINI, permutation, SHAP or the magnitude of coefficients from any regression algorithm? Why or why not? Please discuss in the manuscript.

3.It is not clear from the manuscript whether the feature importance values are derived from the wearable technology. How? Were these variables actually measured ? and are they present in any dataset available in the public domain?

4. You are correct, there is a chance that there is a slight imbalance in class data. I moved to stratified random sampling, with any imbalance in the datasets being a result of a rounding error (Methods).

5. In addition to Pearson, I also included a Gini Importance metric to measure feature importances. Using Gini makes sense here because it is based off of a Random Forest model, which was determined to be one of the highest performing models (Discussion).

6. To clarify, the dataset used for wearable technology is simply a subset of the original dataset. It only includes features that can either be measured by wearable technology, or can easily be user-inputted (Discussion).

7. No, we did not test for multicollinearity in our dataset. We ran a VIF test with a cutoff value of 6. All features with a VIF value above that is removed from the dataset. From this dataset, however, we chose to keep chest pain (cp) because it is an important feature that can also be used with our wearable technology dataset (Methods).

Thank you for addressing my comments. Accept.