A Comparison of Decision Tree and Random Forest on the Myocardial Infarction Dataset

Abilash Nair

Motivation and Description of the Problem

Heart disease is one of the most commonly occurring medical conditions in the modern world due to a myriad of reasons, especially due to the rising stress levels and obesity. It is very challenging to predict the prognosis of the heart condition. In this coursework, two classifiers will be built for multi-class classification.

- Random forest and decision tree algorithms will be used to build the classifiers
- The problem will be limited to determining the likelihood of patient survival three days after reporting the first myocardial infarction incident
- The classifiers will also attempt to predict the most likely cause of heart failure causing death
- Comparing the effectiveness of the models, with neural network classifier performance as benchmarks

Hypothesis

- Random forest model will perform better than the decision tree model for multi-class classification, especially when it comes to predicting the majority class. Decision trees is better for predicting minority classes
- Random forest classifiers will be a lot more computationally resource intensive and will be more time consuming to train as opposed to a decision tree classifier
- Neural network classifiers built for the dataset will perform better than decision tree or random forest classifiers

Model Comparison

Decision Trees:

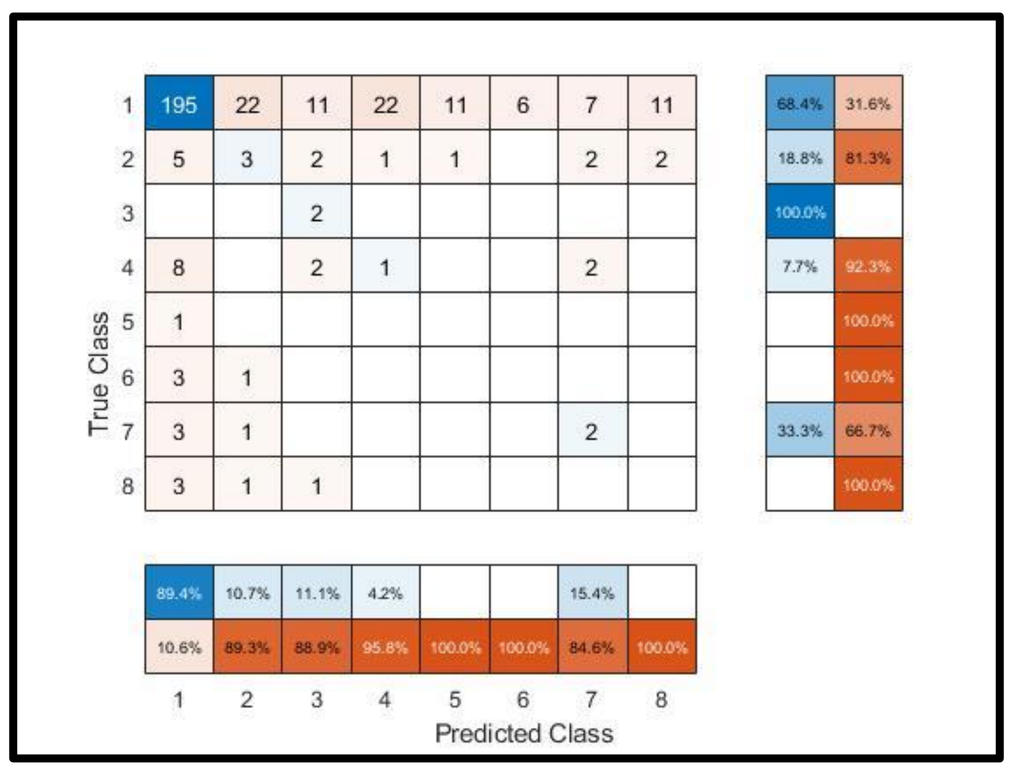
- It is a graphical representation which looks like a tree diagram with nodes and branches that aide in the decision-making process. Decision trees are supervised learning models.
- The classification trees are categorical, for e.g. binary classifications or multi-class classification. Regression trees are more complex as they are used for continuous data types.
- Easier to comprehend the decision trees due to its simple and clear visual structure. The model can be explained well as it is not completely a black box model
- Relatively easier to model computationally, less time and resources needed for training
- Decision trees tend to overfit easily during training and incorrect classifications can be an issue with unseen test data. Variance and bias are issues due to overfitting
- The decision tree could become complex if it creates additional nodes or if the dataset is large, making it difficult to comprehend, more computationally resource intensive, and time intensive

Random forest:

- In a random forest, there are various uncorrelated decision trees that operate together. Whichever outcome is selected by most of the decision trees, that would become the model's prediction, in a classification task.
- The performance and accuracy of a random forest model surpasses that of a single decision tree. This is because errors produced by each of the decision tree offsets each other in a random forest. Individual trees trained on sub samples of the data set are averaged to reach a final decision [1]
- Accuracy of random forest is predicted to be a lot higher than decision trees as decision trees tend to overfit. Errors in individual trees can be offset with other trees in the random forest
- Large datasets can be handled by random forests
- Complex and not very easy to comprehend compared to an individual decision tree
- There could be bias in high values due to the individual decision trees favouring.

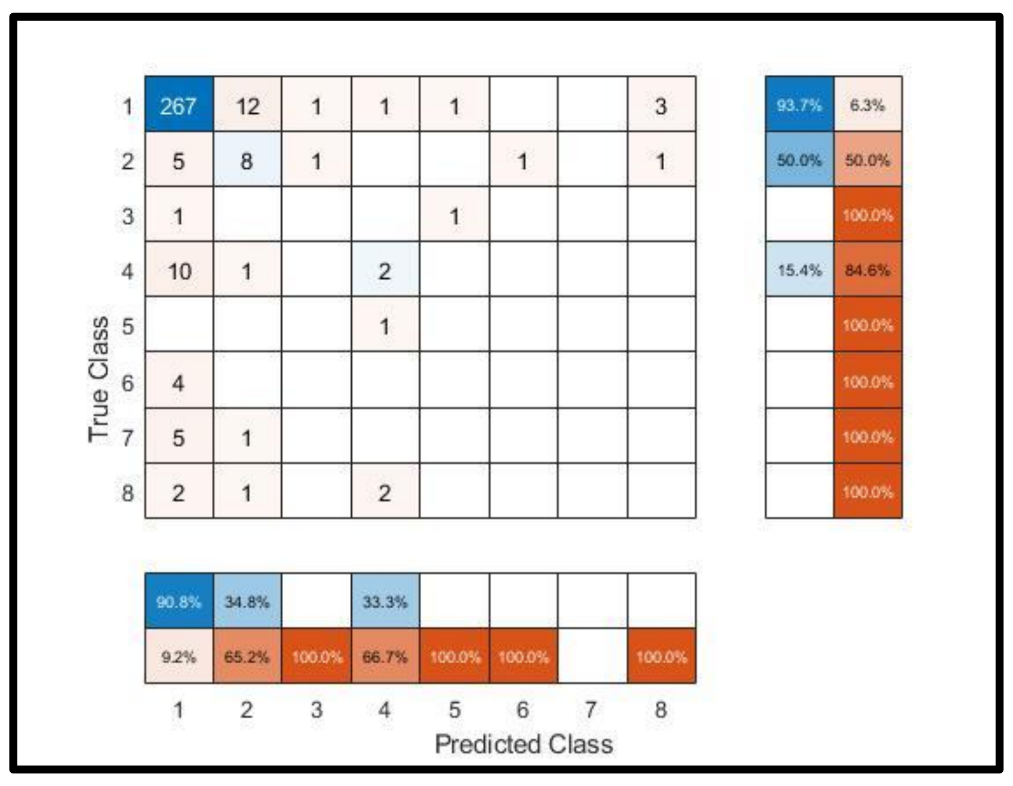
Exploratory Data Analysis

- The myocardial infarction dataset is obtained from UCI repository and contains 104 useable variables, with nine variables that are continuous and the others categorical. Total of 1700 observations.
- Multi-class classification is needed as the label is split into 8 classes, with 1 referring to the patient surviving the MI incident and the others referring to patient dying form the incident. For instance, from cardiogenic shock or pulmonary fibrillation
- Continuous variables are scaled through standard scaler (normalisation)
- Histogram shows a few continuous and categorical variables. The class distribution is also shown. Most of the continuous variables are normally distributed with some showing significant skew. Class distribution is severely imbalanced with majority class being 1 (patient survived)
- Box plots of all continuous variables show the variance and outliers of data distribution
- Serum content, WBC content, and ESR content contain a significant number of outliers.
- Correlation heatmap using Pearson correlation does not provide significant information regarding correlations of categorical data. For continuous data, some variables such as blood pressure and ESR levels show correlation with lethal outcome



	1	2	3	4	5	6	7	8		
1	195	22	11	22	11	6	7	11	68.4%	31.6%
2	5	3	2	1	1		2	2	18.8%	81.3%
3			2						100.0%	
4	8		2	1			2		7.7%	92.3%
5	1									100.0%
6	3	1								100.0%
7	3	1					2		33.3%	66.7%
8	3	1	1							100.0%
	89.4%	10.7%	11.1%	4.2%			15.4%			
	10.6%	89.3%	88.9%	95.8%	100.0%	100.0%	84.6%	100.0%		

Figure 5 – DT Confusion Matrix



	1	2	3	4	5	6	7	8		
1	267	12	1	1	1			3	93.7%	6.3%
2	5	8	1			1		1	50.0%	50.0%
3	1				1					100.0%
4	10	1		2					15.4%	84.6%
5				1						100.0%
6	4									100.0%
7	5	1								100.0%
8	2	1		2						100.0%
	90.8%	34.8%		33.3%						
	9.2%	65.2%	100.0%	66.7%	100.0%	100.0%		100.0%		

Figure 6 – FT Confusion Matrix

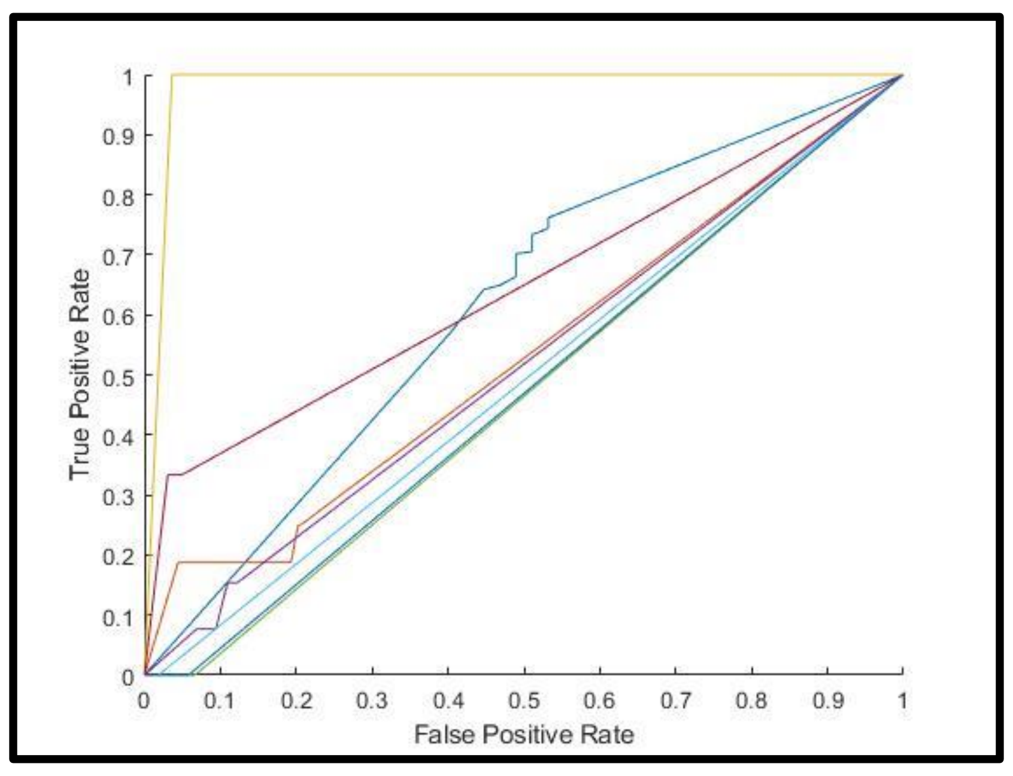


Figure 7 – DT ROC Curve

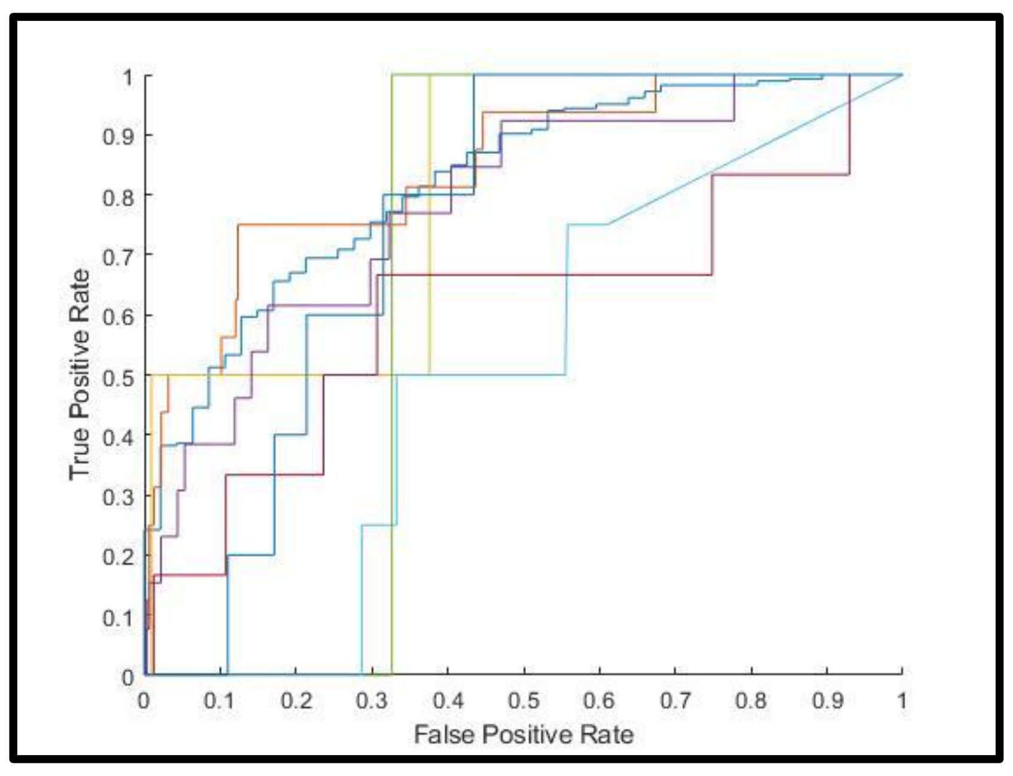


Figure 8 – FT ROC Curve

RF Model Majority Class Preditction Performance								
Class	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
Accuracy	86.45%	93.07%	98.80%	95.48%	99.10%	98.49%	98.19%	97.29%
Precision	90.82%	34.78%	0.00%	33.33%	0.00%	0.00%		0.00%
Recall	93.68%	50.00%	0.00%	15.38%	0.00%	0.00%	0.00%	0.00%
Specificity	42.55%	95.25%	99.39%	98.75%	99.40%	99.70%	100.00%	98.78%
F1Score	92.23%	41.03%		21.05%				

Table 1 – RF Model Metrics

Evaluation Methodology

- Due to severe imbalance, oversampling through SMOTE [2] and under sampling is required to increase the number of training samples of minority classes to an acceptable level. Split into training and testing datasets in a roughly 75% to 25% split
- The hyperparameters in the decision tree classifier is tuned through Bayesian optimisation and then compared with the results from grid search
- The hyperparameters in the random forest classifier is tuned purely through grid search. The Out-of-Bag OOB error will be used to tune the model.
- The critical metrics such as classifier performance measures, error metrics (OOB or MSE error), ROC curves with F1 scores, and time taken will be used to evaluate and compare both the models

Choice of Parameters and Experimental Results

Decision Tree:

- Maximum number of splits and minimum leaf size (depth) are used as parameters for tuning
- Bayesian optimisation is used as the optimiser. Compared with results from grid search for reference
- Bayesian optimiser yielded values of 1 for minimum leaf size and 208 for maximum number of splits, while grid search yielded 1 for minimum leaf size and 369 for maximum number of splits.
- Bayesian optimised values were chosen, and training time was faster at 0.22 seconds

Random Forest:

- Number of trees, number of predictors, and minimum leaf size are used as parameters for tuning
- Grid search was used for hyperparameter optimisation for the three parameters
- 13 best predictors were identified, which had a scale above 1.0 (chosen arbitrarily).
- Parameter values are chosen such that it minimises Out-Of-Bag OOB error. Parameter values that maximise the prediction accuracy of majority classes was also evaluated
- Parameter values minimising OOB error are 15 for the number of parameters, 5 for minimum leaf size, and 70 for number of trees, while values maximising majority class accuracy are 10, 5 and 100 respectively
- The performance was comparative with accuracy at 86.75% for OOB minimised and 88.86% for accuracy maximised. However, the time taken to train is much longer at 58 seconds compared to 38 seconds.

Analysis and Critical Evaluation of Results

- Random forest classifier strongly outperformed decision trees when it came to majority class classification, with an accuracy of 86.45% as opposed to 65.96%. Therefore, random forest classifiers can be used to determine a patient's survivability after the initial myocardial infarction incident
- However, the performance of both the classifiers to accurately predict minority classes are poor. Overall, minority classification remains a challenge for both models, and this could be due to the severe imbalance in the dataset. Even though SMOTE was applied to the dataset, the initial number of samples within minority classes were too low for it to be significantly effective
- Accuracies for the RF model remained in the high 90s for all classes. However, this could be due to the much better predictive ability of the RF classifier to classify the majority class [3]
- Analysing the F1 scores reveal that generally the RF model at 92% performed better than DT at 77.5%. Precision is more important than recall, as the aim of the classifier is to predict the number of people who will not survive
- The ROC for each class in the RF model has generally a good curve especially for the majority class
- The DT model suffer from overfitting of training data. This causes it to poorly generalise for the test data, while the RF model generalises well on the test data for the majority class
- The hyperparameter tuning was much faster for the DT models than the RF, with DT tuning taking around 46 seconds while the RF tuning took nearly 2 hours. This is because the DT tuning is only meant to optimise a single tree while the RF tuning ranged from 50 to 100 trees. Also, the RF tuning optimised three variables as opposed to two for DT.
- Training speed was also much faster for the DT model, which only took 0.22 seconds, as opposed to 38 seconds for the RF model. The RF model hyperparameter tuning as well as model training was much more computationally resource intensive and time consuming than DT models
- For the RF model, the number of trees needed for good results were from 70 to 100. A larger number of trees could aggregate to estimate minority classes better or solidify the majority class prediction
- The number of parameters needed were also few with 15 being the optimum number of parameters. The feature importance plot supports this as only around 13 predictors had a significant impact on classification
- A 2-layer neural networks outperformed the DT and RF classifiers for both majority and minority classification, with an estimated average error of 5% for all predicted classes [4]. A similar paper also shows a neural network achieving in excess of 90 percent accuracy [5]. The RF model built in this coursework can still be a good alternative for neural networks

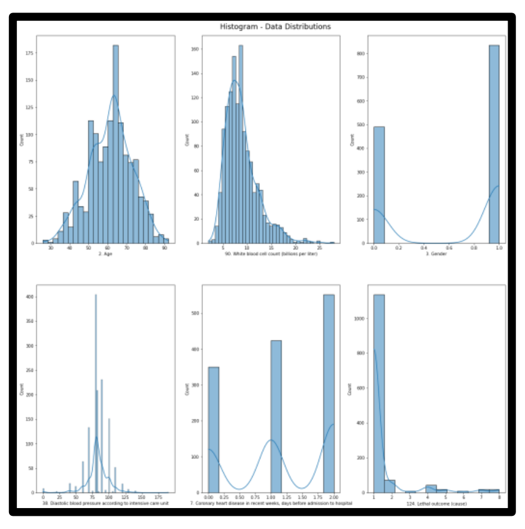


Figure 1 - Histogram

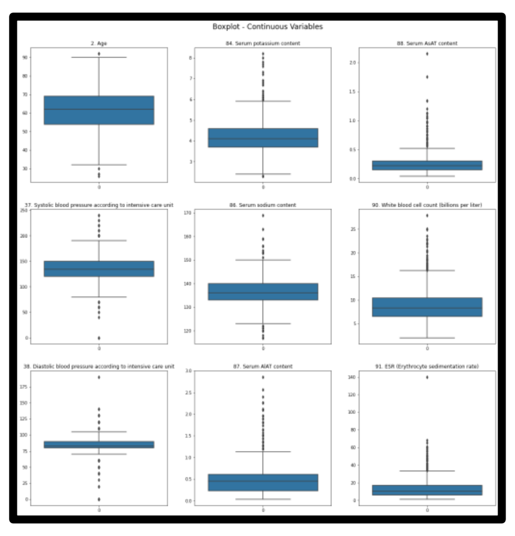


Figure 2 - Box

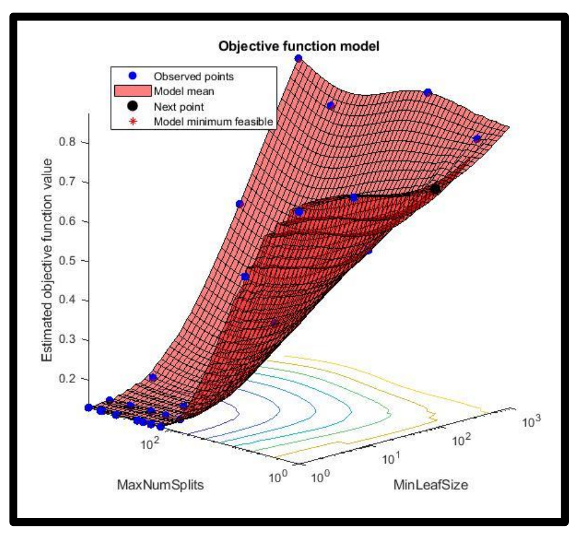


Figure 3 – DT OFM

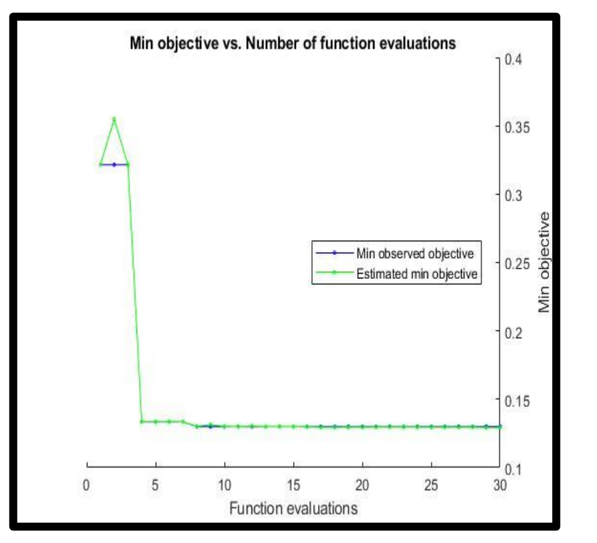


Figure 4 – DT Train

Lessons Learned and Future Work

- Imbalanced data cause significant challenges to training ML models especially when it is coupled with low number of data samples. It might have been better to focus on building a binary classifier that predicts patient survivability than trying to incorporate causes of heart failure leading to death as well
- Although OOB error was used to optimise the RF model, K-fold optimisation could have been used as a benchmark when comparing RF with DT or other models
- Alternate techniques to overcome the limitations of imbalanced data must be explored to avoid misclassifying minority classes
- Feature engineering and clustering techniques can be explored to overcome the limitations of limited samples and ineffective variables. Clustering techniques can also provide a means to visualise the spread of minority classes and the best way to oversample such classes

References

[1] Bishop, C., 2016. *Pattern recognition and machine learning*. New York: Springer.

[2] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp.321-357.

[3] Chen, C., Liaw, A. and Breiman, L., 2004. *Using Random Forest to Learn Imbalanced Data*. [online] Department of Statistics, University of California, Berkeley. Available at: <https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf> [Accessed 15 December 2021].

[4] Dorrer, M., Golovenkin, S., Nikulina, S., Orlova, Y. and Pelipeckaya, E., 2020. Using artificial intelligence to predict the human body's response to cardiovascular disease. *Journal of Physics: Conference Series*, 1679(4), p.042012.

[5] Golovenkin, S., Dorrer, M., Nikulina, S., Orlova, Y. and Pelipeckaya, E., 2020. Evaluation of the effectiveness of using artificial intelligence to predict the response of the human body to cardiovascular diseases. *Journal of Physics: Conference Series*, 1679(4), p.042017.