CS430 Report
Factors that Increase Chances of Obtaining Cardiovascular Disease

u2001509

Abstract

This paper will analyse data relating to cardiovascular disease from across the world. In particular from Cleveland, Hungary, Switzerland, and the VA Long Beach. The paper will then attempt to identify trends to identify factors that increase chances of having a cardiovascular disease. The data used is from *Kaggle* (1), based on data from *UC Irvine* (2).

I. Introduction

Across the world, cardiovascular disease is the leading cause of death. In 2019 alone, 17.9 million deaths were caused by cardiovascular disease (CVD), which accounted for 32 ± 1% of all global deaths (3) (Note: Other sources differ slightly). There is existing research on the dataset that will be considered. Machine Learning techniques are currently the focus of numerous research studies within the healthcare industry. This paper will analyse datasets to determine factors that may contribute towards a person developing CVD. Data analysis techniques will be used in order to detect trends, and determine the contributing factors and the reasons for this.

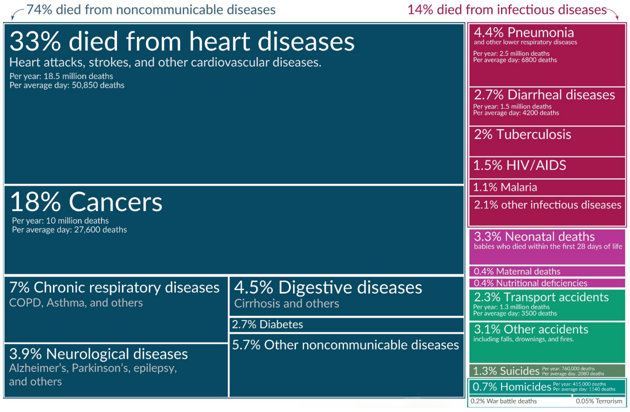


Figure 1: Chart of causes of mortality in 2019 from Our World in Data (4). Heart Diseases account for 33% of all causes of death.

II. Background

A. Machine Learning in Healthcare

Technology is increasingly being used within the healthcare industry. It is a rapidly growing area, with a lot of funding being focused into achieving fast and accurate diagnoses. Many innovations have been made within the healthcare industry using artificial intelligence techniques(5). Some notable examples are image recognition software to detect cancer cells (6), creating and discovering new drugs (7) and machine learning techniques for diagnosing diseases (8).

There have been ethical worries regarding patient data and AI, particularly those using more complex models (9). The main issues are regarding the use of patient data in large datasets and the extent to which they can be used within Machine Learning models, biases or discrimination within models (10) and the 'Black Box Problem' (11). The 'Black Box Problem' relates to issue with *how* Deep Learning models come up with their conclusions. Sometimes it is difficult for researchers to understand how a model arrives at its conclusion, which is a big problem for healthcare in particular as physicians must understand decisions about their patients, both for diagnosis and for recommending treatments (12). For this reason, simpler classification and clustering models are more widespread within healthcare. Examples will be discussed in this report.

B. Cardiovascular Disease

Cardiovascular Disease (CVD) is an umbrella term for a group of disorders of the heart and blood vessels. Any abnormalities in normal blood flow from the heart may result in several types of heart disease. Examples of CVDs include things such as coronary heart disease, stroke and deep vein thrombosis (3). Some of the main factors that increase the risk of getting CVD include but is not limited to smoking, obesity, blood pressure, ethnicity, high cholesterol, diabetes, age and sex (13) (14).

Some of these risk factors are categorical, and some are continuous. This means both classification and clustering techniques can be used to determine correct boundaries for each of these factors in determining risks. The number of contributing factors is also vast and varied so is important to determine which contribute the most towards the risk of developing a CVD. In general, the cause of CVD can be grouped into three categories: behavioral factors, metabolic factors and environmental factors.

III. The Data

The datasets to be used for this investigation are listed below:

A. UC Irvine Machine Learning Repository

UC Irvine is a research-based University in California. In particular, they host a large Machine Learning Repository with several thousand databases. The database used for this analysis contains information regarding cardiovascular disease across four areas in the world (Cleveland, Hungary, Switzerland, and the VA Long Beach). The data contains up to 76 attributes, and all patient data has been anonymised to ensure patient confidentiality. The data used in this analysis is from 1988 and was the most recent data available (2), and has been cited in 64 reports.

In 2019, a report using the data investigated Machine Learning Techniques for predicting cardiovascular disease and determined the accuracy of these techniques using different models (Table 1). The table shows a high accuracy and will attempted to be reproduced within this report.

Model	Accuracy	Classification Error	Precision
Naive Bayes	75.8	24.2	90.5
Logistic Regression	82.9	17.1	89.6
Decision Tree	85.0	15.0	86
Random Forest	86.1	13.9	87.1
SVM	86.1	13.9	86.1

Table 1: Condensed Table from report (15) investigating different Machine Learning techniques on the UC Irvine Machine Learning data for predicting Heart Disease diagnoses as percentages.

B. Kaggle

The original data was transferred to Kaggle (16), for data analysis. In the original dataset, all the data is spread across four different Comma Separated Value (CSV) files relating to the different areas in the world. The dataset has now been combined and condensed. 14 key attributes have been chosen within the dataset, which is also recommended within the original dataset from UC Irvine. There are 5 continuous features and 8 categorical features (along with the target).

Both datasets have been consolidated within another Kaggle dataset and the data has been augmented for further Machine Learning use (1). This is the data that is used for the report. I have reached out to the author of the dataset for clarification on the augmentation method used.

IV. Software and Techniques

The data is given as a Comma Separated Value (CSV) file which was then used with various programs and tools to analyse the data:

A. Microsoft Excel

Microsoft Excel is used to initially examine the data as a CSV. It is useful to use its inbuilt features to overview the initial data and identify any early issues as well as any trends.

B. Weka

Weka is useful particularly for its graphical representation of the data. This is useful to identify trends. It is also used to test different classification and clustering models rapidly. Metrics are given within Weka as well as confusion matrices for model evaluation.

C. SQL

SQL is a query language and is used mainly in exploratory data analysis and data cleaning. When preprocessing data, it is useful for finding null values and anomalies as well as creating helpful sub-tables.

D. Python

Python is used for modelling and feature analysis. After experimenting with classification and clustering in Weka, a more robust model can be achieved within Python. It is also useful for creating plots and graphics within the report.

E. R

R is used for statistical analysis. It assists in fitting statistical distributions to data, as well as clustering and classification. R can be used for data augmentation and can create models.

V. Hypothesis

Several factors contribute to heart disease. Features such as age, diabetes, prior heart disease, and blood pressure have been shown to reflect the most important risk factors for heart disease in previous investigations (17). Additionally, a high accuracy is expected, particularly for classification methods as seen in similar studies (15); however we may expect lower precision than desired, especially for a medical environment.

Classification models may perform better than clustering because in diagnosing CVD, because they reflect more closely to a process doctors would follow. Different risk factors such as blood pressure and BMI, may result in patients being classified into different risk strata of developing CVD.

VI. Data Cleaning and Preprocessing

Before looking at trends, the data must be cleaned to remove anomalous data and identify missing data.

A. Missing Values and Duplicates

After examining the data in SQL and Python, there were no missing values or duplicate records found. Ordinarily there are agreed techniques for handling missing values such as replacing them with the mean value for that category or discarding the incomplete record in further data analyses.

B. Outliers

To ensure correlations could correctly be drawn from the data and to ensure the data was accurately acquired, it was decided to investigate outlying points which lay more than 3 standard deviations from the mean. There were 2009 such datapoints that fell under this definition of an outlier. There were some clearly anomalous values, for example one record where a person who was 59 years old had a weight of 11kg and height of 1.79m, resulting in a BMI of 3.5kgm^{-2}. Another record had a BMI that is nearly 300, but the largest ever recorded BMI is 105. These and many other records were very improbable so these particular rows were discarded so the data was not skewed. The outliers were not substituted with the mean value or another metric because there were numerous outliers which may have misrepresented the data. They were likely made by human error. 3 standard deviations was chosen because using 1.5 * IQR would result in the removal of many valid datapoints. 4 and 5 standard deviations from the mean kept too many anomalous values in the data, such as a weight of 11kg, or a BMI of over 75kgm^{-2}.

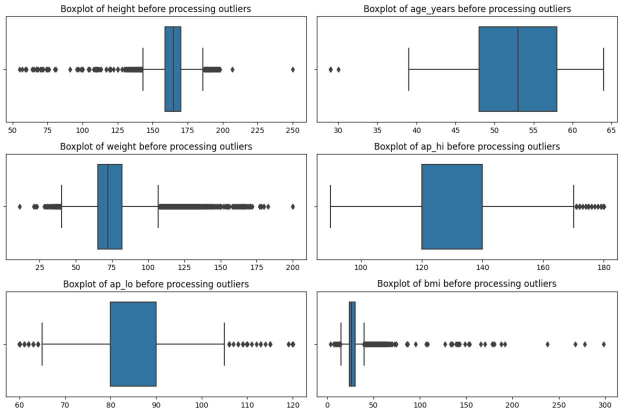


Figure 2: Boxplots of numerical data before removing outliers

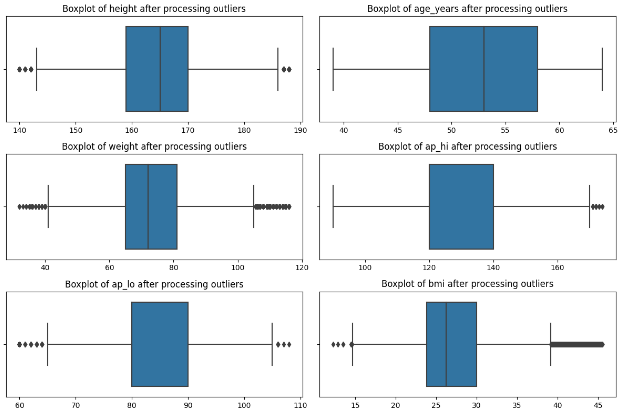


Figure 3: Boxplots of numerical data after removing outliers

After removing outliers, there were 65, 000 records which is enough to conduct the investigation. The IQR and mean values were similar before and after removing outliers.

C. One Hot Encoding

Within the dataset, some of the features are categorical. Some of these are binary such as ‘gender’, ‘alco’ and ‘smoke’ (which indicate whether somebody drinks alcohol or smokes respectively). Others such as ‘cholesterol’ or ‘gluc’ (glucose) have three or four different categories. In total there are 8 categorical features, as well as the target feature. For some machine learning techniques in Python, it is useful to ‘One Hot Encode’ the non-binary categorical data columns. This is a technique used to represent categorical variables as binary vectors. It involves creating a binary column for each unique category, assigning a value of 1 to the column corresponding to the category present, and 0 to others, enabling the representation of categorical data in a format suitable for numerical computation.

D. Scaling

Some models are sensitive to the scales of the data, especially if some features have different magnitudes to others. This is especially important in distance based models such as ‘K-Nearest-Neighbours’ and most clustering algorithms. For this reason a second dataset that uses the ‘Standard Scaler’ in the Python package *sklearn* is used to standardise the data. It uses the formula $\frac{x-\mu}{\sigma}$, where x is the datapoint in a column, and μ, σ are the mean and standard deviation respectively. This is done on both numerical and categorical data.

VII. Data Analysis

A. Exploratory Data Analysis (EDA)

To find any biases in data collection and distribution of variables, univariate analysis was conducted. This is shown in Figure 4. Within the figure, we can see that not all the data is distributed uniformly. Within the binary data, the data for alcohol consumption and smoking is skewed. This could lead to later bias in the models such as all those who smoke being labelled as developing CVD. This is important to be aware of when looking through the data. Those with more than one category such as glucose levels have similar issues. Other variables such as gender and the target variable of having a CVD are distributed more uniformly.

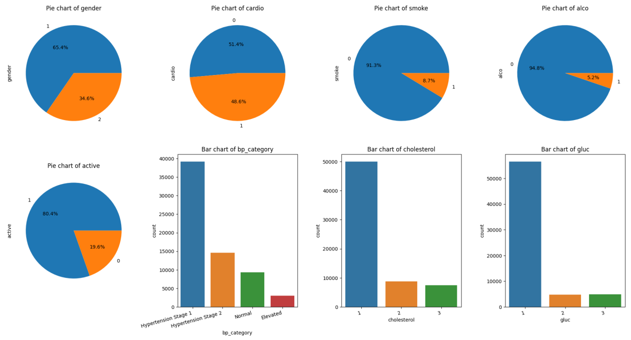


Figure 4: Pie charts and bar charts to show univariate distributions.

Numerical data was also investigated, which showed both age as roughly uniform, and blood pressure following a normal distribution.

The class imbalance could be fixed by augmenting the data using techniques in R such as ‘Smote’. Ensemble algorithms such as Gradient Boosting and Random Forests deal with imbalanced classes well so may perform better for this reason. Techniques such as cross validation may also be used to get a model fit when the data is imbalanced.

B. *Feature Importance*

Feature Analysis and importance shows us which variables contribute to the target variable the most. It is also important to investigate the underlying correlations within the data. This is useful to see which variables are related, whether positively or negatively. In Figure 5, there is a correlation matrix showing all correlations between the variables. Correlation, $\rho_{X,Y}$ is defined as:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(X)}} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

This gives a value in the interval $[-1, 1]$, where 1 means there is a strong positive correlation between variables, and -1 means there is a strong negative correlation. The closer the value to 0, the smaller the correlation. Clearly, we can see a strong positive correlation between variables such as systolic and diastolic blood pressure, or weight and BMI which is expected. There are some weaker positive correlations such as that between glucose and cholesterol levels. Blood pressure, cholesterol levels and age are the most correlated with developing a CVD.

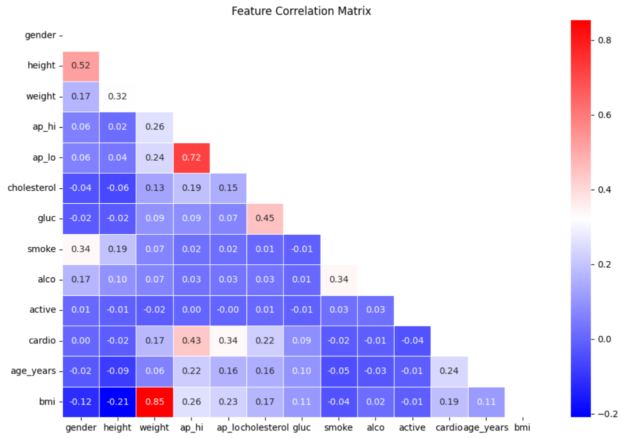


Figure 5: Correlation map of the variables. Red squares are strongly correlated, blue squares are not correlated. There is no negative correlation.

A more robust way to discover feature importance is by running a Machine Learning model. After running a simple random forest in Python, we can see how the model prioritises some variables relative to others. As shown in Figure 6, the three variables that contribute the most to classifying if someone has a CVD are BMI, age and blood pressure. Alcohol consumption, smoking and activity levels contribute the least for this particular model, which may not be as expected. To determine how well models run, we need to investigate metrics of how well the different models perform.

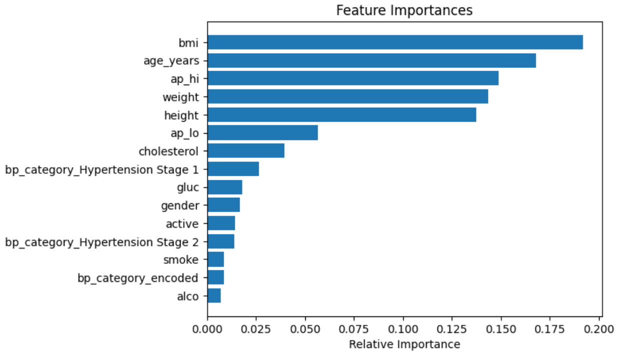


Figure 6: Feature importance determined from a random forest model.

VIII. Modelling

Classification models can be made within both Weka and Python. Results are shown within Table 2 with various error metrics calculated for each model, calculated from confusion matrices. The models were trained and tested with a 66% split. With some accuracies over 70%, this represents a reasonable fit, implying that the models can predict the chances of someone developing a CVD reasonably well.

Model	ROC AUC	Accuracy	Precision	Sensitivity
Naive Bayes	77.4	70.6	74.6	60.2
Logistic Regression	79.3	72.8	76.2	64.5
Decision Tree	63.1	63.1	62.4	61.1
Random Forest	76.5	70.1	70.1	69.0
SVM	78.8	71.2	78.6	57.6
Gradient Boosting	80.3	73.5	75.5	67.7
KNN (N=5)	72.9	68.3	68.6	64.6

Table 2: Error Metrics over various models as percentages

The first metric that is used is:

$$\text{ROC AUC} = \int_0^1 \text{TPR}\, d(\text{FPR})$$

This can be interpreted as the area under the ROC (Receiver Operating Characteristic) graph such as in Figure 7. It is a graph of True Positive Rate (TPR $= \frac{\text{TP}}{(\text{TP+FN})}$) against False Positive Rate (FPR $= \frac{\text{FP}}{(\text{FP+TN})}$). The area under the curve (ROC AUC) ranges from 0.5 to 1, with an output of 0.5 showing that class allocation are all due to chance. With a value here of 0.8, it can be ascertained that our model has allocated classes well.

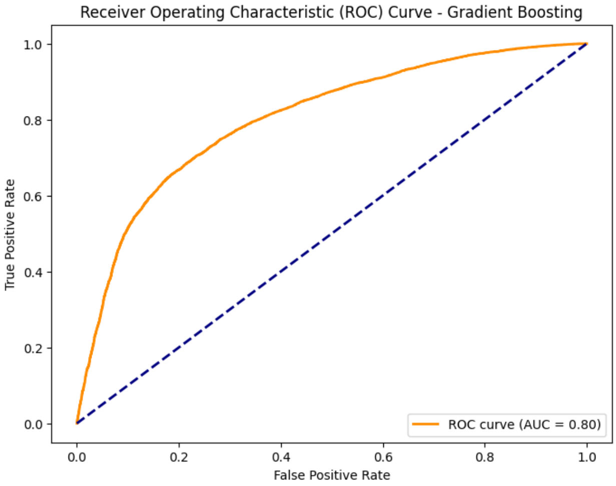


Figure 7: ROC Curve for Gradient Boosting

The other metrics used are:

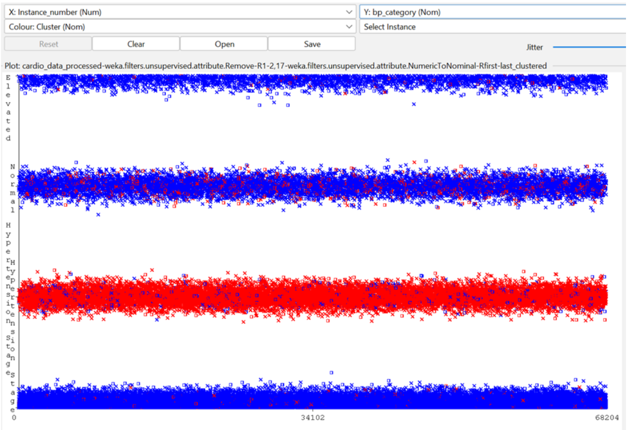
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

Where TP, TN, FP and FN are True Positives, True Negative, False Positives and False Negatives respectively.

Several clustering algorithms were also used within WEKA, such as SimpleKMeans and DBSCAN. These attained lower accuracies than classification models. SimpleKMeans resulted in an accuracy of 63.7%. There may be several reasons as to why clustering performed worse than classification. However one reason could be due to scaling. The values that were one-hot encoded may require further scaling to reduce distances. Clustering algorithms are very sensitive to distances as that is how the clusters are determined. If categorical data is given values of 0 and 1, these are relatively far apart in contrast to other numerical values. An attempt was made to scale these categorical values during preprocessing; however the scaling may not be suitable. When investigating clustering diagrams, Blood Pressure was the feature that assigned clusters the most as show in in Figure 8. Many other features were not good at distinguishing between the two.


X: Instance_number (Num)

Y: bp_category (Nom)

Colour: Cluster (Nom)

Plot: cardio_data_processed-weka.filters.unsupervised.attribute.Remove-R1-2,17-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last_clustered

Figure 8: Cluster assignment on blood pressure types.

Compared to the feature importance in classification modelling, the results are very different. Blood pressure was the only distinguishing factor in choosing clusters. As none of the other features contributed, this is likely to be the cause of the poor performance in clustering.

Finally, all classification models were run on the original data, before preprocessing it. The aim of this is to see the effect of preprocessing the data and whether it had a positive or negative impact on the results. This was done on the same seed as Table 2 to make the best comparison. As can be seen in Table 3, the results show the preprocessing had a positive impact on the results. For ensemble based algorithms such as Random Forest, Decision Tree and Gradient Boosting, the results did not change much. The accuracies did increase slightly in some cases where the data was unprocessed. The reason for this, as mentioned earlier, is because these models handle unbalanced class distributions well. This is the same case for Naive Bayes. Others, such as Logistic Regression, SVM and KNN had accuracies that decrease drastically (over 13% in the case of KNN). This is a good sign that the preprocessing was done well especially for class imbalances and evaluation of outliers.

Model	ROC AUC	Accuracy
Naive Bayes	77.4	70.7
Logistic Regression	75.6	69.6
Decision Tree	63.3	63.4
Random Forest	77.7	71.7
SVM	63.4	59.4
Gradient Boosting	79.5	72.8
KNN (N=5)	56.9	55.2

Table 3: Error Metrics for original data over various models as percentages

IX. Conclusions

From Table 2, we have seen that our models predict classification of having a CVD reasonably well. However our results are still not good enough for use in medical practice. The best accuracy was with a Gradient Boosting model, with an accuracy of 73.5%. Overall, the best model was also the Gradient Boosting Model as it has the best accuracy and ROC AUC, and the second best in precision and sensitivity.

Within a medical environment, precision and sensitivity is crucial. Precision can be interpreted as the proportion of true positive values compared to false positive values. False positives could worry a patient and can lead to unnecessary follow-up tests, treatments, or interventions. This can result in additional stress, cost, and potential harm to the patient especially as being diagnosed with a CVD is worrying. Similarly, sensitivity is the rate of true positives to false negatives. Again, we want to minimise false negatives as they can result in a failure to detect a condition that requires treatment. This may lead to delayed or missed interventions, potentially allowing the condition to progress and cause harm to the patient.

Clearly from Table 2, our precision and sensitivity values are much lower than desired for a medical environment. These values should be as close to 100% as possible. Additionally, our accuracies are not high enough to just rely upon the models. In practice, the results indicate that the models do not perform well when viewed in isolation, but may be used to indicate or help a physician in predicting CVD.

Compared to published models in Table 1, which used the same dataset, the values using the models in this analysis were worse. This is likely due to differences in preprocessing and outlier evaluation, as well as subtle differences in the models used. Within the report (15), deep learning and SVM techniques achieved the best results, with Gradient Boosting only performing better than Naive Bayes. In the results in Table 2, almost the opposite was found, showing the importance of choices in preprocessing.

The models could be improved using various techniques. Earlier, it was noted that there may be biases within the data when conducting univariate analysis. A dataset with a more uniform distribution may improve the accuracy of the dataset and limit underlying bias. Alternatively, the data could be augmented using various techniques to achieve the same result, although real data may be more accurate. The data was also taken from only four locations. It is possible that obtaining data from more varied sources could increase the accuracy. Similarly, only 14 key attributes were chosen out of a possible 76. Although computationally expensive, it may improve accuracy to consider other features that could contribute better towards predictive models.

X. Extensions

Possible extensions are as follows:

1. The dataset used did not include many environmental or location based data. This is due to the fact if we split the data into location, there would not have been enough data to investigate the trends. An alternative dataset can be used to investigate the impact of ethnicity and environmental location in developing a CVD (18).

2. As CVD is an umbrella term for numerous diseases, more investigation could be made classifying specific types of CVD. This would have to be incorporated within the dataset if using the models from this report.

3. There were some issues with the dataset, so it may be better to investigate alternative or supplementary datasets. The main issue within this dataset is the large number of categorical features. For example, activity levels are given on a scale of 1 – 4 rather than how many hours of exercise there are per week. This also applies to other risk factors e.g. cholesterol levels. If the data given were numerical rather than categorical, a better model may have been made. The data was also quite old. More recent data may result in slightly different trends and may be more reliable data as patient data is digitally stored automatically rather than taken by hand and transcribed.

4. Due to the scope of this report, a handful of models were selected. Future reports may investigate other models to find a model with greater accuracy and precision. If a complex model is used, further analyses could investigate how ethical the models are. This is a requirement within healthcare and a current worry about machine learning (19).

XI. Bibliography

[1] Cardiovascular Disease — kaggle.com, `https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease/data`, [Accessed 21-12-2023].

[2] M. P. R. D. A. Janosi, W. Steinbrunn, UCI Machine Learning Repository — archive.ics.uci.edu, `https://archive.ics.uci.edu/dataset/45/heart+disease`, [Accessed 18-12-2023].

[3] Cardiovascular diseases (CVDs) — `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, [Accessed 17-12-2023] (Jun. 2021).

[4] S. Dattani, F. Spooner, H. Ritchie, M. Roser, Causes of death, Our World in DataHttps://ourworldindata.org/causes-of-death (2023).

[5] How Artificial Intelligence is Accelerating Innovation in Healthcare — goldmansachs.com, `https://www.goldmansachs.com/intelligence/pages/how-artificial-intelligence-is-accelerating-innovation-in-healthcare.html`, [Accessed 18-12-2023].

[6] D.-M. Koh, N. Papanikolaou, U. Bick, R. Illing, C. E. Kahn Jr, J. Kalpathi-Cramer, C. Matos, L. Martí-Bonmatí, A. Miles, S. K. Mun, et al., Artificial intelligence and machine learning in cancer imaging, Communications Medicine 2 (1) (2022) 133.

[7] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, M. J. Ahsan, Machine learning in drug discovery: a review, Artificial Intelligence Review 55 (3) (2022) 1947–1999.

[8] M. M. Ahsan, S. A. Luna, Z. Siddique, Machine-learning-based disease diagnosis: A comprehensive review, in: Healthcare, Vol. 10, MDPI, 2022, p. 541.

[9] M. D. McCradden, J. A. Anderson, E. A. Stephenson, E. Drysdale, L. Erdman, A. Goldenberg, R. Zlotnik Shaul, A research ethics framework for the clinical translation of healthcare machine learning, The American Journal of Bioethics 22 (5) (2022) 8–22.

[10] B. Giovanola, S. Tiribelli, Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms, AI & society 38 (2) (2023) 549–563.

[11] J. J. Wadden, Defining the undefinable: the black box problem in healthcare artificial intelligence, Journal of Medical Ethics 48 (10) (2022) 764–768.

[12] B. Chan, Black-box assisted medical decisions: Ai power vs. ethical physician care, Medicine, Health Care and Philosophy (2023) 1–8.

[13] E. C. Leritz, R. E. McGlinchey, I. Kellison, J. L. Rudolph, W. P. Milberg, Cardiovascular disease risk factors and cognition in the elderly, Current cardiovascular risk reports 5 (2011) 407–412.

[14] Cardiovascular disease — nhs.uk, `https://www.nhs.uk/conditions/cardiovascular-disease/`, [Accessed 19-12-2023].

[15] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE access 7 (2019) 81542–81554.

[16] M. Y. H, Heart Disease Dataset — kaggle.com, `https://www.kaggle.com/datasets/yasserh/heart-disease-dataset`, [Accessed 18-12-2023].

[17] M. S. Pathan, A. Nag, M. M. Pathan, S. Dev, Analyzing the impact of feature selection on the accuracy of heart disease prediction, Healthcare Analytics 2 (2022) 100060.

[18] T. E. O'Toole, D. J. Conklin, A. Bhatnagar, Environmental risk factors for heart disease, Reviews on environmental health 23 (3) (2008) 167–202.

[19] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, M. Ghassemi, Ethical machine learning in healthcare, Annual review of biomedical data science 4 (2021) 123–144.