Application of Distributed Computing and Neural Networks for Heart Disease Detection

John Doyle   
MSc in Data Analytics *CCT Dublin*Cork, Ireland  
sba23437@student.cct.ie

Heart disease is one of the largest fatal diseases health care practitioners (HCPs) face in modern society. It effects developed and under-developed countries and is increasing globally in incidence. Heart disease occurs where the heart fails to supply sufficient blood throughout the body. Early diagnosis of the disease helps mitigate damage caused by the lack of blood flow and saves lives ultimately. For HCPs, there are medical diagnoses available such as angiographies, but techniques like this are invasive and costly, especially in terms of HCP resourcing and in terms of access for patients. Often an angiography is scheduled following patient complaints, meaning significant disease impact may already have occurred or the patient may be in later stages of the disease.

In response to this data, scientists and health researchers are teaming up to leverage patient data as an alternative approach for diagnosing disease in patients. This approach is of course non-evasive, and the data needed in many of the prediction algorithms are easily recorded/ made available such as heart rate monitoring, blood pressure monitoring and self-reported pain.

In this study, an intelligent predictive system is implemented for the identification and diagnosis of cardiac disease building on former studies. Irrelevant and noisy data is removed using advanced techniques. The classification rate of the developed system is examined. It is anticipated that the enhanced predictive system will have potential for application in future HCP settings to help diagnose heart disease early, accurately, and effectively.

Keywords—ANN, Machine Learning, Heart Disease, HDFS, PySpark,

# Introduction & Motivation

Heart disease is a significant global cause of mortality, affecting many individuals. Early identification of heart disease is crucial, as it can save lives. A powerful tool in the field of medical diagnostics is machine learning (ML), which includes techniques like Neural Networks (Muhammad, Tahir, Hayat, & Chong, 2020, p8). Many health initiatives require the prevention of the disease and early identification of diseases. Getting an intervention in place early really puts patience in a strong position when dealing with diseases like heart disease. Often patients report having pain in their chests which then leads to diagnosis of the condition, but earlier detection means the pain may be off set for longer increasing their quality of life. But in the absence of chest pain, what indicators can be leveraged? We have a wide and deep breadth of data available in current times from smart devices and home monitoring systems to self-reported data. Pulling on these modern data sources including data such as heart rate or blood pressure or reported chest pain is a convenient and useful method to predict health problems (Ismail, Abdlerazek, & El-Henawy, 2020, pp15–19).

## Big Data and Hidden Patterns

Special characteristics that led to the popularity of big data are referred to as the 3Vs of volume, velocity, and variety. Each year, the quantity of data generated online rapidly increases, so big data applications benefit decision-makers by identifying correlations, enabling the review of massive datasets, spotting trends, and presenting data clearly to others (Liao, Tang & Luo, 2018, p166).

The health sector generates vast amounts of data. This data can be structured, unstructured, or semi-structured. Big Data in recent times has been leveraged to understand User/ Device and Log Data in the Healthcare. There are many examples of Bio-medical data with the potential to provide predictive knowledge of pathological features for biomedicine. (Tang et al., 2021, P5). Big data analytics allows us to uncover hidden information and intricate patterns that might not be apparent through traditional clinical analysis. By applying ML techniques, we can process and analyse this data effectively (Ismail, Abdlerazek, & El-Henawy, 2020 pp15-19). Accurate forecasting and decision assistance may be achieved in an effective manner with machine learning (ML). Big Data, or the vast amounts of data generated by the health sector, may enable models to make diagnostic choices by revealing hidden information or intricate patterns (Rao et al, 2024).

## Neural Networks

In recent years, Neural Networks (NN) have seen widespread and successful implementation in a wide range of data applications, often surpassing other classifiers (Altaf, Soomro, & Rawi, 2019, pp59-64). Specifically, researchers have developed deep learning algorithms specifically for heart disease detection. These algorithms combine different neural network architectures and other techniques to enhance accuracy. Recursive feature elimination (RFE) for example is helpful for identifying the most key features for prediction models (ScikitLearn, 2024)

In summary, Neural Networks, when combined with other techniques (RFE), play a crucial role in accurate heart disease detection. Their ability to learn from complex data patterns makes them valuable tools for improving patient outcomes and saving lives.

# Methodology

Bottom-up approaches to health care are becoming increasingly prevalent in health research due to

• the huge growth and availability of patient health information,

• the systems supporting health care and patient health care management,

• and with the development of machine learning options.

Exploring the raw health data collected by patients and / or health care providers can provide us a method for identifying patterns in various scenarios. For example, take the patient’s data set and analyse the results to see if the doctors need to diagnose the patient for disease diagnosis/ treatment. There are three major steps in Big Data processing,

1. Data collection/ Raw Big Data,

2. data processing (preprocessing, data representation/ encoding/ normalisation), feature selection to create predictive knowledge,

3. and prediction improvement using AI such as NNs to create predictive results (Tang et al., 2021, P6).

There are different systems for doing this type of work and Hadoop Distributed File System (HDFS) for example supports large data processing efficiently (Tang et al., 2021, P6). Big data analytics refers to innovative analytic approaches scaled to enormous datasets from terabytes (TB) to zettabytes (ZB) of various types, such as structured, unstructured, and semi-structured data1,2. Big data analytics can be used on datasets that vary in size compared to traditional databases with few capabilities to capture processes and manage the data patterns (Rao et al, 2024). Based on that, I am leveraging Pyspark on Hadoop in the preprocessing and simulation phases of the data pipeline. One of the key features Spark offers for speed is the ability to run computations in memory, but the system is more efficient than MapReduce for example for complex applications running on disk (Karau, Konwinski, Wendell, & Zaharia, 2015, P1).

In this study, I am leveraging Machine Learning models to provide large data analysis for heart disease detection. Using Apache Hadoop as the development platform, the suggested framework for heart disease prediction is displayed in Figure. 1.

A diagram of a flowchart

Description automatically generated

1. framework for heart disease prediction.

An Artificial neural network (ANN) is a system inspired by our central nervous systems whereby neurons fire or are activated based on inputs or connections to other neurons. If a certain threshold is hit or a certain array are activated it creates an output. ANNs use this same idea to decide or process complex arrangements of data inputs and can create models or flows of information through neurons arranged in layers. ANNs are particularly good with fitting problems, with enough neurons ANNs can fit most data. Neural Network links a set of input nodes existing in the input layer with a set of one or more output nodes existing in the output layer through an intermediate hidden layer. In recent studies, ANNs have been shown to perform well predicting outcomes on large datasets with multi-variate relational variables (Khaldi, El Afia, Chiheb & Faizi, 2017, pp1-6).

In this study, I procured the health data that had the potential to predict patient outcomes as it relates to heart disease. More on the data in the next section. I used supervised learning in this study to train the model based on patient features and the known outcomes for each patient. In this study, I will leverage similar approaches to previous studies such as (Altaf, Soomro, & Rawi, 2019, pp59-64), where researchers utilized supervised ANN architecture for data prediction efficiency showcasing their outcomes with a confusion matrix.

The neural network proposed for this research have a multi-layer neural network operating under supervised learning; it consists of two layers including one input layer, and one output layer. I had the option to create additional hidden layers, but more complex builds are not needed where a simple model can perform effectively and accurately. It has been described in other research that, increasing the number of hidden neurons does not necessarily enhance the learning results, but could only increase the process of learning time (Brouwer, 1997, pp.117–126). In this experiment the number of neurons and layers were fixed through trial and error.

# DATA SET & PREDICTOR PREPARATION

For the purposes of this assignment, I have procured the “Heart Disease” database from the UC Irvine Machine Learning Repository (1998). In many studies, reuse of codes and datasets in ANN development is common, with public repositories like GitHub being extremely helpful for researchers and developers. It is valuable for researchers to analyse the same datasets using different techniques to replicate and validate findings or to lend support to novel approaches in modern machine learning (Ghofrani, Kozegar, Bozorgmehr, & Divband Soorati, 2019, p8).

The selected data consists of thirteen features and one target as outlined in Appendix A.

The data contains health related data and outcomes for heart disease for 270 patients made available to the public for the purposes of researching heart disease. In the data preparation for this study, I took steps to check for data types, checked the validity of the data, examined the shape and descriptive statistics of the data (Figure 2), recoded categorical data to integer type data for use in the ANN in later stages and determined there were no outliers and null values.

A table with numbers and letters

Description automatically generated

1. Descritive Statistics

The dataset was next pre-processed using standardization (Figure 3) which helps the ANN perform better and removes the potential for large numbers/ small numbers to have a bias on the model outcomes.

A white rectangular box with black text

Description automatically generated

1. Normalizing The Data

After normalising the data, I checked what sort of relationships existed between all the features and the dependent variable or the heart disease outcome in other words. Sometimes in studies in this area, multiple features have the potential of co-corelating with the outcome. This has the potential to bias the ANN towards these features which seem to all partially similar effects. This might give those features too much predictive importance. It also means we may be processing too many features/ redundant features which adds to the model processing time and cost. In the OLS Regression Results (Figure 4) we can see most features have a significant probability of having a correlational relationship with the dependent variable (P>0.05). However, a few features, in particular ‘Age’ do not seem to be significant. The model also provides a warning that there appears to be strong multicollinearity of our variables. This is visually demonstrated in image five where we can see patterns that are common in the correlations between the different variables. This is something we can use in the next steps for our feature selection.

A screenshot of a computer

Description automatically generated

1. Regression Results

## Recursive Feature Elimination (With Cross Validation)

Steps normally for long numerical data include impute missing data, outlier removal, then normalise the numerical data and lastly feature selection is done generally with correlational analysis (Tang et al., 2021, pp.13-14). Recursive feature elimination (RFE) was next used to identify the most key features (See Figure 5). keeping non-informative features leads to over-fitting and is therefore detrimental for the statistical performance of the models. Some machine learning algorithms can be misled by irrelevant input features, resulting in worse predictive performance. (Scikitlearn, 2024).

A screenshot of a computer program

Description automatically generated

1. RFE

A screenshot of a computer

Description automatically generated

1. RFE Ranking

In a second step to this cross validation was used to help identify features that may be ineffective or redundant for this analysis. Age in particular, was identified using this method as not been in the mix of optimal features (Figure 6).

Finally, the dataset was split into training (80%) and testing data (20%) to validate the model’s predictive accuracy (See Figure 7).



1. Train Versus Test Split

# Results and Discussions

I created a layered sequential model (Figure 8), with five hundred neurons in the first layer, using rectified linear activation function or ReLU for short. ReLu is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance. On the other hand, for binary classification (where there are two possible classes), sigmoid activation function is the appropriate choice and I have leveraged it in the output layer and indicated the output has a single neuron for the ‘0’ or ‘1’ output value (Brownlee 2020).

For tuning the parameters, my assumption is that the relationship between the dataset and the predicted health outcome may not be overly complex which implies one layer is sufficient to analyse the relationships between the features as predictors. Regarding the neurons, I leveraged fifty, then one hundred and finally five hundred in the single hidden layer with five hundred providing the best prediction outcome.

Across many of the different studies included in the review, it seemed widespread practice to generally leverage 100 to 150 epochs with batch sizes of ten. I kept this standard for our data set. For larger datasets with more complexity, it is worth tuning these ten aspects to increase performance, but increasing batch size also costs resources and processing time which must be considered as part of the overall calculation.

A computer screen shot of text

Description automatically generated



1. Model Configuration

The model performed well, completing in a brief time (4 ms/dtep) with an accuracy of 100% and loss of 0.5% (See Figure 9). It took just 10 epochs to reach an accuracy of greater than 90%, but with a loss of about 24%. Up to epoch eighty the model slowly gets to 100% accuracy while reducing the loss to about 3%. By epoch 120 the loss is down to 1%. The benefit of a lower loss value in our learning model is like Validity and Reliability in general research terms. Accuracy could be mapped to validity, whereas loss could be mapped to reliability. Will the model provide a valid prediction each time it runs, and will our model reliably provide a similar prediction given certain inputs each time? Both are needed for us to have confidence in the model’s predictive ability. Getting 100% accuracy and a loss 0.5% are good indicators for this prediction model.

A screenshot of a computer

Description automatically generated

1. Evaluate the Model

Iterating through the first few instances in the training dataset (Figure 10), we see a perfectly predicted outcome for the first five rows in image eleven.

A screenshot of a computer code

Description automatically generated

1. First 5 Instances of Testing

Finally, the model was evaluated using the test data set (20% of the overall original dataset) with excellent outcomes. In the confusion matrix (Figure 10) we can see no false positives or false negatives and a 100% accuracy of prediction using our trained model with our test data. Using this approach randomly splitting the dataset into a train and test set aligns with the approaches used by similar other studies investigated in the review of the literature and is the minimal testing required to provide some level of confidence in a new model. Following from this, we will need to acquire an updated dataset with more recent patient data and apply the model to compare its ability to predict. We will want to see predictive power that is not biased by the first dataset/ the first data collection methodologies, the first data set feature definition or the first dataset publication bias/ omitted data. If the data and features are truly controlled, I expect the model will perform excellently on future outcome predictions.

A screenshot of a computer

Description automatically generated

1. Confusion Matrix

# Concluding Remarks

The template is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

Many of the studies examined in preparation for this study describe a limitation around the reproducibility or generalisability of developed learning models. After concluding this investigation, there is value in keeping approaches as streamlined and as straight forward as possible. This way researchers have logical generalisability for new implementations, mapping similar datasets to existing pipelines or saved models. However, this is probably the limit to how generalisable any saved model is due to the potentially unlimited variability of data sets and processed features. This open nature of discovery lends itself to discovery and innovation and determines future pathways towards best practice and evidence-based practices in machine learning for health care. The major conclusion from this research is how accessible and available solutions are for guided decision-making in health scenarios. The major recommendation following from this is regarding our data repositories. Having excellent data governance, data definition, feature governance and feature definitions will be key where there is a desire to have increased generalisability of developer learning models. The governance and definitions must be scientific and reduce bias,

• from how data is measured,

• from how data is reported

• and from missing data

These will be key ingredients in future Big Healthcare Data.

##### References

1. Altaf, S., Soomro, M. W., & Rawi, M. (2019). Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining. ICISDM 2019: Proceedings of the 2019 3rd International Conference on Information System and Data Mining, 59-64. doi: 10.1145/3325917.3325919
2. Brouwer, R. K. (1997). Training a feed-forward network by feeding gradients forward rather than by back-propagation of errors, Neurocomputing, 16, 117–126
3. Brownlee, J. (2020). A Gentle Introduction to the Rectified Linear Unit (ReLU). https://MachineLearningMastery.com, [Accessed 20 March 2024]
4. Ghofrani, J., Kozegar, E., Bozorgmehr, A., & Divband Soorati, M. (2019). Reusability in Artificial Neural Networks: An Empirical Study. In 23rd International Systems and Software Product Line Conference - Volume B (SPLC ’19), September 9–13, 2019, Paris, France. ACM, New York, NY, USA, Article 4, 8 pages. DOI: 10.1145/3307630.3342419
5. Ismail, A., Abdlerazek, S., & El-Henawy, I. M. (2020). Big data analytics in heart diseases prediction, Journal of Theoretical and Applied Information Technology, 98(11), 15–19
6. Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). Learning Spark: Lightning-Fast Big Data Analysis. O’Reilly Media, Inc.
7. Khaldi, R., El Afia, A., Chiheb, R., & Faizi, R. (2017). Artificial Neural Network Based Approach for Blood Demand Forecasting: Fez Transfusion Blood Center Case Study, 1-6. DOI: 10.1145/3090354.3090415
8. Liao, H., Tang, M., Luo, L., et al. (2018). A bibliometric analysis and visualization of medical big data research. Sustainability, 10(1), 166
9. Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific Reports, 10, 19747
10. Rao, G. M., Ramesh, D., Sharma, V., et al. (2024). AttGRU-HMSI: Enhancing heart disease diagnosis using a hybrid deep learning approach. Scientific Reports, 14, 7833. DOI: 10.1038/s41598-024-56931-4
11. Recursive Feature Elimination (RFE) for Feature Selection in Python (2024). scikit-learn 1.4.1 documentation. Retrieved from https://machinelearningmastery.com/rfe-feature-selection-in-python/, [Accessed 20 March 2024]
12. Recursive feature elimination with cross-validation. (2024). scikit-learn 1.4.1 documentation. Retrieved from https://scikit-learn.org/stable/auto\_examples/feature\_selection/plot\_rfe\_with\_cross\_validation.html#sphx-glr-auto-examples-feature-selection-plot-rfe-with-cross-validation-py, [Accessed 20 March 2024]
13. Tang, L., Li, J., Du, H., Li, L., Wu, J., & Wang, S. (2021). Big Data Analytics in Forecasting Research: A Literature Review. Big Data Research, 27, 100289
14. UCC Irvine Machine Learning Repository (1998), Retrieved from https://archive.ics.uci.edu/dataset/45/heart+disease , [Accessed 20 March 2024]
15. Data Source

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Variable Name** | **Role** | **Type** | **Description** |
| 1 | Age | Feature | Integer | In years |
| 2 | Sex | Feature | Categorical | (1 = male; 0 = female) |
| 3 | Chest pain type | Feature | Categorical | chest pain type  -- Value 1: typical angina  -- Value 2: atypical angina  -- Value 3: non-anginal pain  -- Value 4: asymptomatic |
| 4 | BP | Feature | Integer | resting blood pressure (on admission to the hospital) mm Hg |
| 5 | Cholesterol | Feature | Integer | serum cholesterol mg/dl |
| 6 | FBS over 120 | Feature | Categorical | fasting blood sugar > 120 mg/dl |
| 7 | EKG results | Feature | Categorical | resting electrocardiographic results  -- Value 0: normal  -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)  -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria |
| 8 | Max HR | Feature | Integer | maximum heart rate achieved |
| 9 | Exercise angina | Feature | Categorical | exercise induced angina (1 = yes; 0 = no) |
| 10 | ST depression | Feature | Integer | ST depression induced by exercise relative to rest |
| 11 | Slope of ST | Feature | Categorical | the slope of the peak exercise ST segment  -- Value 1: upsloping  -- Value 2: flat  -- Value 3: downsloping |
| 12 | Number of vessels fluro | Feature | Integer | number of major vessels (0-3) coloured by fluoroscopy |
| 13 | Thallium | Feature | Categorical | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| 14 | Heart Disease | Target | Integer | diagnosis of heart disease (angiographic disease status)  -- Value 0: < 50% diameter narrowing  -- Value 1: > 50% diameter narrowing |