**Task (4000-6000 words) Harvard Referencing**

● Overview of the chosen topic, including objective statement and Research Question. Presentation of state of the art, including research methodologies and key of the papers you reviewed. [0 - 20]

● Literature review (10 References Min). [0 - 15]

● Critical evaluation of the key findings, specifically their implications and limitations, and highlighting any contradicting viewpoints and research gaps. [0 - 30]

● Conclusions you have drawn based on your research. [0 - 15]

**Title**

Application of Distributed Computing and Neural Networks for Heart Disease Detection

**Abstract**

**Introduction and 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. One powerful tool in the field of medical diagnostics is machine learning (ML), which includes techniques like Neural Networks (Muhammad, Y., Tahir, M., Hayat, M. et al.).

Most health education initiatives require the prevention of the disease and early identification of diseases [1]. Big data analysis in healthcare is very convenient and useful to use technology to produce medical data with spark and machine learning algorithms to predict health problems [2] (Ismail, A., Abdlerazek, S. & El-Henawy, I. M, 2020)

Big Data and Hidden Patterns

The health sector generates vast amounts of data, often referred to as big data. This data can be structured, unstructured, or semi-structured. 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, A., Abdlerazek, S. & El-Henawy, I. M).

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 assist models used to make diagnostic choices by revealing hidden information or intricate patterns (Rao, G.M., Ramesh, D., Sharma, V. et al, 2024).

Big Data in recent times has been leveraged to understand User/ Device and Log Data mostly. There are many examples of Bio-medical data with the potential to provide predictive knowledge of pathological features for biomedicine. (Tang, Ling & Li, Jieyi & Du, Hongchuan & Li, Ling & Wu, Jun & Wang, Shouyang. (2021, P5).

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 visualizations benefit decision-makers by identifying correlations, enabling the review of massive datasets, spotting trends, and presenting data clearly to others. Big data visualization techniques incorporate presentation methods for any type of data in a graphical format, which eases interpretation and understanding (Liao, H. et al., 2018).

Neural Network Approach

In recent years, Neural Network (NN) has seen widespread and successful implementations in a wide range of data mining applications, often surpassing other classifiers (Student Performance Prediction using Multi-Layers Artificial Neural Networks A Case Study on Educational Data Mining (1))

Researchers have developed hybrid 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) helps identify the most important features for disease prediction ([Recursive Feature Elimination (RFE) for Feature Selection in Python - MachineLearningMastery.com](https://machinelearningmastery.com/rfe-feature-selection-in-python/))

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 is becoming more and more 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. Take the patient’s data set and get the results to see if the doctors need to diagnose the patient.

There are 3 major steps generally in Big Data processing; Data collection/ Raw Big Data, 2; data processing (preprocessing, data representation/ encoding/ normalisation), feature selection to create predictive knowledge and lastly 3; prediction improvement using AI such as NNs to create predictive results (Tang, Ling & Li, Jieyi & Du, Hongchuan & Li, Ling & Wu, Jun & Wang, Shouyang. (2021, P6).

Hadoop Distributed File System (HDFS) supports large data processing efficiently Tang, Ling & Li, Jieyi & Du, Hongchuan & Li, Ling & Wu, Jun & Wang, Shouyang. (2021). 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 (Rao, G.M., Ramesh, D., Sharma, V. et al, 2024).

I am leveraging Pyspark on the Hadoop data in the preprosessing and simulation phases of the data pipeline. One of the main features Spark offers for speed is the ability to run computations in memory, but the system is more efficient than MapReduce for complex applications running on disk (Holden Karau, Andy Konwinski, Patrick Wendell, Matei 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 **Fig. 1**.

**Fig. 1**

A diagram of a flowchart

Description automatically generated

Artificial neural networks (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 very 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).

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.

Second, the user should build and design his network by choosing the type of learning: supervised learning, unsupervised learning or reinforcement learning. As well as by fixing the network parameters, for example net input, transfer function, learning function, learning rate, number of neurons in each layer, etc. In this study, I will leverage similar approaches to previous studies such as (Altaf, Saud & Soomro, Muhammad Waseem & Rawi, Mohd. (2019)), 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 2 layers including one input layer, and one output layer. I had the option to create additional hidden layers, but more complex builds aren’t needed where a simple model can perform effectively and accurately. It’s been described in other research that, increasing the number of hidden neurons don’t neccesisaroly enhance the learning results, but could only increase the process of learning time (Brouwer, R. K. (1997) 117–126.). In this experiment the number of neurons and layers were fixed through trial and error.

**DATA SET AND PREDICTOR PREPARATION**

For the purposes of this assignment, I’ve procured the ‘Heart Disease’ database from the UC Irvine Machine Learning Repository (Irvine 1998, [Heart Disease - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/45/heart+disease)). 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’s valuable for researchers to analyse the same datasets using different techniques in order to replicate and validate findings or to lend support to novel approaches in modern machine learning (Javad Ghofrani, Ehsan Kozegar, Arezoo Bozorgmehr, and Mohammad Divband Soorati. 2019).

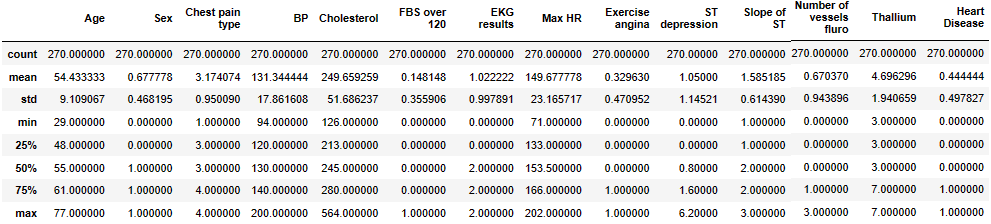
The selected data consists of 13 features and 1 target as follows:

Table1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **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 |

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 (Table 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.

Table2



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

Image 3

A white rectangular box with black text

Description automatically generated

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 (Image4) 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 5 where we can see patterns that seem to be common in the correlations between the different variables. This is something we can use in the next steps for our feature selection.

Image 4

A screenshot of a computer

Description automatically generated

Image 5

A screenshot of a computer

Description automatically generated

Recursive Feature Elimination (With Cross Validation)

Steps normally for loog numerical data include impute missing data, outlier removal , then normalise the numerical data and lastly feature selection is done generally with correlational analysis (Tang, Ling & Li, Jieyi & Du, Hongchuan & Li, Ling & Wu, Jun & Wang, Shouyang. (2021, p13-14). Recursive feature elimination (RFE) was next used to identify the most important features (See Image 6). 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).

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 (Image7).

Image 6

A screenshot of a computer program

Description automatically generated

Image 7

A screenshot of a computer

Description automatically generated

Finally, the dataset was split into training (80%) and testing data (20%) to validate the model’s predictive accuracy (See image 8).

Image 8



**Results and Discussions**

I have created a layer sequential model (Image 9), with 500 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 indicated the output has a single neuron for the ‘0’ or ‘1’ output value (Brownlee 2020).

Image 9

A computer screen shot of text

Description automatically generated

The model performed well, completing in a short time (4 ms/dtep)with an accuracy of 100% and loss of 0.5% (See image 10)

Image 10

A screenshot of a computer

Description automatically generated

Iterating through the first few istnaces in the training dataset, we see a perfectly predicted outcome for the first 5 rows in image 11.

Image 11

A screenshot of a computer code

Description automatically generated

Finally, the model was tested using the test data (20% of the overall original dataset) with excellent outcomes. In the confusion matrix (Image 11) we can see no false positives or false negatives and a 100% accuracy of prediction using our trained model with our test data.

Image 11

A screenshot of a computer

Description automatically generated

**References**

1. Javad Ghofrani, Ehsan Kozegar, Arezoo Bozorgmehr, and Mohammad Divband Soorati. 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. https://doi.org/10.1145/3307630.3342419
2. Artificial Neural Network Based Approach for Blood Demand Forecasting: Fez Transfusion Blood Center Case Study (Khaldi, El Afia, Chiheb & Faizi) 2017
3. Altaf, Saud & Soomro, Muhammad Waseem & Rawi, Mohd. (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. 10.1145/3325917.3325919.
4. Tang, Ling & Li, Jieyi & Du, Hongchuan & Li, Ling & Wu, Jun & Wang, Shouyang. (2021). Big Data Analytics in Forecasting Research: A Literature Review. Big Data Research. 27. 100289. 10.1016/j.bdr.2021.100289.
5. Learning Spark, Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, Inc., 2015.
6. [Recursive feature elimination with cross-validation — scikit-learn 1.4.1 documentation](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)
7. Rao, G.M., Ramesh, D., Sharma, V. et al. AttGRU-HMSI: enhancing heart disease diagnosis using hybrid deep learning approach. Sci Rep 14, 7833 (2024). <https://doi.org/10.1038/s41598-024-56931-4>
8. Liao, H. *et al.* A bibliometric analysis and visualization of medical big data research. *Sustainability* **10**(1), 166 (2018).
9. BIG DATA ANALYTICS IN HEART DISEASES PREDICTION, Ismail, A., Abdlerazek, S. & El-Henawy, I. M. Big data analytics in heart diseases prediction. *J. Theor. Appl. Inf. Technol.* **98**(11), 15–19 (2020). [BIG-DATA-ANALYTICS-IN-HEART-DISEASES-PREDICTION.pdf (researchgate.net)](https://www.researchgate.net/profile/Ahmed-Ebada-2/publication/342349215_BIG_DATA_ANALYTICS_IN_HEART_DISEASES_PREDICTION/links/5eef9bb3a6fdcc73be911e42/BIG-DATA-ANALYTICS-IN-HEART-DISEASES-PREDICTION.pdf)
10. Recursive Feature Elimination (RFE) for Feature Selection in Python, [Recursive Feature Elimination (RFE) for Feature Selection in Python - MachineLearningMastery.com](https://machinelearningmastery.com/rfe-feature-selection-in-python/)
11. Early and accurate detection and diagnosis of heart disease using intelligent computational model, Muhammad, Y., Tahir, M., Hayat, M. *et al.* Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci Rep* **10**, 19747 (2020). <https://doi.org/10.1038/s41598-020-76635-9>
12. Brouwer, R. K. (1997). Training a feed-forward network by feeding gradients forward rather than by back-propagation of errors. Neurocomputing, 16, 117–126.
13. (Brownlee 2020) A Gentle Introduction to the Rectified Linear Unit (ReLU) [A Gentle Introduction to the Rectified Linear Unit (ReLU) - MachineLearningMastery.com](https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/)