**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.)

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)

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**

Data mining is

a method of identifying fascinating patterns in current data in various scenarios to turn the data into valuable

information. Take the patient’s data set and get the results to see if the doctors need to diagnose the patient. Tis

work employs a hybrid deep learning model to provide large data analysis and visualization techniques for heart

disease detection. Using Apache Hadoop as the development platform, the suggested framework for heart disease

prediction is displayed in Fig. 1. An enhanced k-means clustering (IKC) method removes outliers before analyzing the curated medical data. Recursive feature elimination (RFE) is then used to identify the most important

features once the distribution classes have been balanced using the synthetic minority over-sampling method

(SMOTE). Ultimately, the bio-inspired hybrid mutation-based swarm intelligence (HMSI) model employs an

attention-based gated recurrent unit network (AttGRU)to forecast diseases.

A diagram of a flowchart

Description automatically generated

Artificial neural networks (ANN) is an intelligent system inspired by the human nervous system. ANNs are very good with fitting problems, with enough neurons ANNs can fit any data with arbitrary accuracy. Neural Network links a set of input nodes 𝑥𝑖 = (𝑥1, 𝑥2, … , 𝑥𝑛) existing in the input layer with a set of one or more output nodes 𝑦𝑗 = (𝑦1, 𝑦2, … , 𝑦𝑚) existing in the output layer through an intermediate hidden layer. Nodes in each layer are activated once they reach the layer threshold value 𝜃𝑖 . This matching is realized by finding an unknown function h.

yj = h(x1, x2, … , xn) (1) 𝑖 ∈ [1, 𝑛] and 𝑗 ∈ [1, 𝑚]

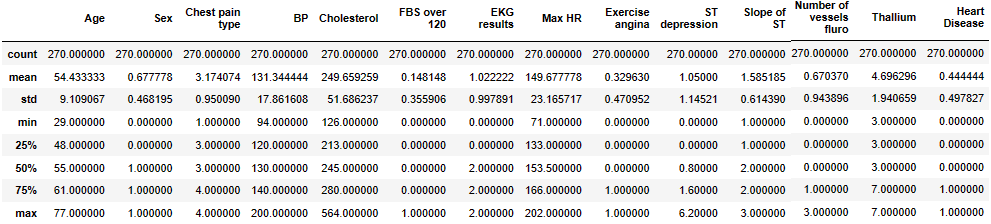
To implement artificial neural networks the user should follow the plotted flow chart in Figure 1. First, data should be collected and partitioned into input dataset and desired output dataset. Second, the user should build and design his network by choosing the type of learning: supervised learning, 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. Then, the dataset should be preprocessed using either normalization or standardization. After, it should be divided on training data, validation data and testing data. In fact, the training dataset is used to identify the values of weights and biases of the network. While the validation dataset is employed to analyze weights and biases, so as to measure the capacity of network generalization, and to interrupt training when overfitting occurs [15]. Whereas, the testing dataset is used to validate weights and biases participating in the stopping criterion, and to evaluate the network performance on new datasets [6]. Finally, the network should be simulated, and if necessary its settings should be modified until obtaining good results.

The neural networks proposed for this research are multilayer feedforward neural networks operating under supervised learning; they consist of three layers including one input layer, one hidden layer and one output layer. Indeed more hidden neurons provide better learning results. However, some previous studies assert that increasing the number of hidden neurons couldn’t enhance the learning results, but could only increase the process of learning time [12]. In this study, hidden neurons were fixed through trial and error, therefore once the number of hidden neurons exceeds 15 the improvements become marginal to null. Hence, the structure of the three models is 10-15-1. Also, ANN parameters were chosen based on trial and errors. In this study, the learning rate is 0.01 and the momentum is 0.001.

**DATA SETS 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)). The data consists of 13 features and 1 target as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Variable Name** | **Role** | **Type** | **Description** |
| 1 | age | Feature | Integer | In years |
| 2 | sex | Feature | Categorical | (1 = male; 0 = female) |
| 3 | cp | Feature | Categorical | chest pain type  -- Value 1: typical angina  -- Value 2: atypical angina  -- Value 3: non-anginal pain  -- Value 4: asymptomatic |
| 4 | trestbps | Feature | Integer | resting blood pressure (on admission to the hospital) mm Hg |
| 5 | chol | Feature | Integer | serum cholesterol mg/dl |
| 6 | fbs | Feature | Categorical | fasting blood sugar > 120 mg/dl |
| 7 | restecg | 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 | thalach | Feature | Integer | maximum heart rate achieved |
| 9 | exang | Feature | Categorical | exercise induced angina (1 = yes; 0 = no) |
| 10 | oldpeak | Feature | Integer | ST depression induced by exercise relative to rest |
| 11 | slope | Feature | Categorical | the slope of the peak exercise ST segment  -- Value 1: upsloping  -- Value 2: flat  -- Value 3: downsloping |
| 12 | ca | Feature | Integer | number of major vessels (0-3) coloured by fluoroscopy |
| 13 | thal | Feature | Categorical | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| 14 | num | Target | Integer | diagnosis of heart disease (angiographic disease status)  -- Value 0: < 50% diameter narrowing  -- Value 1: > 50% diameter narrowing |



**Results and Discussions**

**References**

**PDF ideas:**

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

* Reuse of codes and datasets in ANN development is common, with public repositories like GitHub being popular.
* Challenges include complexity in finding and making modules reusable.
* Ad-hoc development processes are prevalent in ANN projects.
* Deep Convolutional Neural Networks (CNNs) are more commonly reused.
* Frameworks like TensorFlow and CNTK aid in ANN development.

Artificial Neural Network Based Approach for Blood Demand Forecasting: Fez Transfusion Blood Center Case Study (Khaldi, El Afia, Chiheb & Faizi) 2017

* Blood Demand Forecasting: Utilizing Artificial Neural Networks (ANNs) to predict monthly demand for red blood cells, plasma, and platelets.
* Model Performance: ANN models outperform ARIMA in blood demand prediction.
* Model Building: Three networks are used to forecast each blood component separately.
* Data Handling: Data is partitioned, preprocessed, and distributed for training, validation, and testing.
* Network Structure: Each model has a 10-15-1 structure with specific parameters.
* Transfer Functions: Tangent sigmoid for hidden layers and linear for output layers are employed.
* Applications: Enhancing blood supply chain management and reducing wastage.

Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining

* Equation (1): Defines the relationship between variables in the context of student performance prediction.
* Neural Network Training: Utilizes supervised ANN architecture for data prediction efficiency.
* Predictors: Includes 10 predictors like CourseID, assessments, and grades for forecasting student course outcomes.
* Results: Showcased confusion matrix for testing different architectures with reasonable accuracy rates.
* Future Development: Focuses on extending research to complex data sets and comparing faculties using AI techniques.
* This research aims to enhance educational data mining for student performance prediction using neural networks.

Summary: CNN vs. RNN vs. ANN in Deep Learning

* **CNN:**
  + Ideal for image and video processing.
  + Uses filters to extract features.
  + Captures spatial features and follows parameter sharing.
* **RNN:**
  + Suitable for time series, text, and audio data.
  + Captures sequential information with parameter sharing.
  + Faces challenges with vanishing and exploding gradients.
* **ANN:**
  + Universal function approximators.
  + Suitable for tabular, image, and text data.
  + Loses spatial features but can solve various problems.

Big Data in Forecasting Research: A Literature Review 2021

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 biomendicine. (P5).

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 (P6)

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 (p13-14)

Big Data using Hadoop Framework - Article (1)

Data Processing:

* Big Data platform with Hadoop, MapReduce, and Mahout frameworks handle large datasets for retinal image analysis.
* Hadoop Distributed File System (HDFS) supports large data processing efficiently.

A review of machine learning and big data applications in addressing ecosystem service research gaps

Learning Spark, Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, Inc., 2015.

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 (P1)

Recursive feature elimination with cross-validation

[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)

keeping non-informative features leads to over-fitting and is therefore detrimental for the statistical performance of the models.

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>

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.

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 data3,4.

Liao, H. *et al.* A bibliometric analysis and visualization of medical big data research. *Sustainability* **10**(1), 166 (2018).

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 (5):

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)

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]

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/)

Feature selection refers to techniques that select a subset of the most relevant features (columns) for a dataset. Fewer features can allow machine learning algorithms to run more efficiently (less space or time complexity) and be more effective. Some machine learning algorithms can be misled by irrelevant input features, resulting in worse predictive performance.

Intro deets:

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