USING DECISION TREE AND WEKA TO DIAGNOSE HEART DISEASES FOR ELDERLY PATIENTS

***Abstract*—** The objective of this research paper is to apply K-nn data mining tool on a Thing-Worx platform in order to monitor and predict heart related diseases and detect early vital signs leading to life threatening situations that might lead to stroke, heart attack and angina. Applying that mining technique that can predict heart diseases can assist medical practitioners and health professionals to detect vital signs that leads to critical medical situations. Decision tree, naïve Bayes, support vector machine and neutral network are amongst the classification techniques that have previously been used to address this heart diseases phenomenon.

***Keywords*—** *Heart diseases, Data mining techniques, Machine learning. Algorithms and thing works, (k-nn) K-neighbor-nearest Introduction*

1. INTRODUCTION

Heart disease is amongst the major cause of death throughout the world from the past decade [1] Diabetics, Blood pressure and cholesterol levels are amongst the main contributing factor for heart related diseases. Some of the risk factors associated with heart diseases can be traced back to family medical history, smoking and drinking patterns of the patient and poor health diet. If heart disease vital signs are not properly detected, monitored and contained at an early stage they can lead to critical health conditions such as heart attack, stroke and angina.

A number of data mining techniques such as Decision tree, neural network, Naïve Bayes and K-nn algorithm are been used to predict and detect heart diseases risk and critical vital signs. Furthermore, motivated by the worldwide growth in number of heart diseases related deaths yearly and the availability of patients dataset that has been extracted from various heath care centres, researcher have adapt utilising data mining techniques in order to assist medical care practitioners to diagnose early signs of heart disease [2]

Heart disease prediction and detection systems can help healthcare professionals in detecting and predicting heart diseases based on clinical patient data. Therefore in this research paper, a supervised machine learning J48,Logistic model tree and forest random algorithm together with weka tool is proposed for the implementation of heart disease detection, prediction and monitoring system which will accurately attempt to validate dataset from Cleveland open source. The second contribution in this research paper is the representation of heart disease vital signs monitoring systems that uses IoT concept through sensor, gateway, and Arduino microcontroller. Different medical sensors are been used in the IoT and related technologies in order to collect, analyse and push forward relevant medical information from on node to the other. The thingWorx platform can accommodate a huge number of sensor datasets coming from heterogeneous devices trying to access thingworx server for informed and accurate patient data.

2. Heart diseases and risks factors

2.1 heart disease

A heart is a very sensitive part of a human body which ought to be consciously be guided according to health and medical experts (Aldo et al, 2010). Heart diseases can be defined as any kind of discomfort affecting the normal function of the heart. Heat disease normally occurs whenever the oxygen that is been provided by arteries together with the blood into the heart blocks completely or becomes narrowed ([3] Heart based diseases can be classified in to 4 different categories i.e.

1. **Heart attack**
   Heart attack situation implies data the blood circulation in to the heart is significantly reduced or stopped. It normally happens when one or more coronary arteries which carries the blood to the heart gets blocked. This might lead to cases of the heart muscles getting damaged.
2. **Heart failure**
   This type of situation occurs when the heart can't pump enough supply of blood to various parts of the body .this conditions doesn't automatically mean that the heart has size to function.
3. **Agina**
   This occurs when there is no adequate blood flowing into the heart. It is regarded as amongst the symptoms related to coronary artery diseases, the symptoms ranges from pans in the chests, jaws and shoulders
4. **Arrhythmias**
   These is how slow or fast does the heart beat occurs. It is normally referred to as the irregular manner at which the heart beats per second.

2.2 RISK FACTORS CONTRIBUTING TO HEART DISEASES

The conditions or habits at which an individual is more likely to develop heart related diseases are determine by certain risk factors, which also contribute to the probability of existing disease escalating to a worse state [4]

2.2.1 Controllable risks

Smoking: The traces of chemicals found in the tobacco contributes to the development of blood clots which ultimately increases the heart attack cause due to the plaque that is build-up inside artery walls.

Weight: for every kilograms the body, the highest the risk of diseases. This is medical proven for individuals who carry an extra amount of fat around their waist.

Cholesterol: extreme cholesterol found in the blood, which is building up on the arteries wall can be the cause of atherosclerosis.

Diabetics: Sugar diabetics if it's not properly manage can creates some injury and damages in the artery walls which eventually causes clots within the blood**.**

Blood pressure: The force created by blood against the blood vessels wall can be a contributing factor for blood pressure, normally generated during the pumping of blood within the heart. Whenever the arteries get under an increase amount of pressure a situation called hypertension gets created that further leads to the damage of the arteries walls, coronary and atherosclerosis.

2.2.2 Uncontrollable Risk factors

Age: Most of the heart related disease normally happens when women are have enter the menopause stage and men have past the age of 40. The majority of individuals that die from heart related diseases e.g. heart attack are usually above the age of 65.

Sex: medical studies are of the opinion that men are regarded as having more heart attack risk as compare to women , and most men generally experiences heart attack at an early years of the manhood**.**

Family history: In most instance an individual who have people within his immediate family members who suffered from heart attack is likely to be at risk of heart diseases.

In this research paper we will monitor controllable risk factors such as blood pressure, cholesterol level, chest pains fasting blood, resting blood and fasting blood.

3. MACHINE LEARNING TECHNIQUES

The machine learning techniques for data mining falls under the domain of artificial intelligence (AI), its aims is to:

- construct algorithms that can learn from experiences
- Make accurate predictions for new data se, in this way ML becomes intelligent theory of learning.
- Identifies patterns that are very hard or impossible for human to detect themselves
- It can operate with large dataset and decisions and predictions

Machine learning algorithms used during data mining in most instances can be classified into three categories:

Supervised Learning	Unsupervised learning	Re-enforcement learning
In this type of learning datasets exists with examples Algorithms can learn from the dataset It can respond to new input of what it has learn	Dataset doesn't control the respond in the technique. Algorithms will attempt to recognise any similarity patterns between inputs and classify them based on similarity	This technique falls in between supervised and unsupervised learning. Whereby the model performance improvement occurs as the technique interact with the environment.

Table 1. Difference between supervised, unsupervised and re-enforcement learning.

This research paper will apply supervised machine learning algorithm.

3.1 TECHNIQUES FOR DATA MINING

Data mining technique are nothing else but employed methods that can be utilised for data softening with the aim of identifying any patterns the technique includes amongst its elements of associate , predictions, classifications and clustering.

A. **Associate**: Looks at any form of patterns that is formed based on connection from one event to another. In most cases it is applied in predicting heart diseases whenever the patient of similar risk is identified and further gives the relations that contains unlike feature the space to be analysed [5].

B. Clustering: This techniques is based on unsupervised machine learning whereby there is no class label provided. It identify factors that are previous not know and virtualised them in a collected documented manner**.**

C. Classification: This consists of a supervised learning approach**,** whereby samples of class label training are supplied. Mathematical methods such as k-nn, decision tree , Naïve Bayes and artificial neural network are used for classification (Podgoreclec et al , 2002)

D. **Predication (forecasting**) – This methods is used to determine data patterns which can formulate a future reasonable prediction. This technique can be associated with prognostic model within data mining [6].

E. **Sequence**: This methods what it does is to search for similar pattern whereby one occurrence influence the next future occurrence.

Our research study will focus on classification methods as it relates well to structure learning machine to predict and monitor heart related disease.

3.2 ALGORITHMS FOR DATA MINING

- **Bayesian classifiers**

Through the use of Bayesian classifier, the systems can discover the knowledge that has been concealed that normally can be associated with the diseases based on patient's historical medical records that suffers from heart disease. [7] and [8]. It predicts the class of membership probabilities using Bayes theorem.

- **Decision tree**

This is regarded as the most popular classifier due to its simplicity and easy to implement. It doesn't require parameter setting or domain knowledge[9] and [10].

- **Neutral networks**

This technique is well known based on its results accuracy. It uses a feed forward neural network model, back propagated learning algorithm and variable learning rate [11].

- **Support vector machine**

It is a supervised machine learning which apply state of the art maximum classification algorithm that is deeply rooted in statistical learning theory. It's a technique for both non-linear and linear data classification. The algorithm do searches to identify best hyper planes that can be used to separate the transformed data into two different classes [12].

- **K-neighbour-nearest**

K nearest neighbour (KNN) is simple algorithm, which stores all cases and classify new cases based on the similarities measured. It is classified into two types [12]:

- Structure less NN technique
- Structured based NN technique

KNN is normally used in various applications such as:

- Classification interpretations
- Problem solving
- Function learning
- Teaching and learning

In this research paper we adopt decision tree techniques as we deep it relevant to predict and monitor heart diseases.

3.2 TOOLS FOR DATA MINNING

Rapid miner

Is an open core model software platform that is used to provide data mining tasks that are fully integrated. Besides been a tool for data mining it also has the capabilities allow additional plugins for data analysis algorithms creation.

Weka

Which stands for (Waikato environment for knowledge Analysis) is basically an open source platform that is been development in Java and issued under General public licence (GNU). Fundamentally it is a collection of machine learning algorithms that can be used for data mining tasks such as virtualisation, classification, clustering, data pre-processing and regression**.** Furthermore it has the ability to process data results through the provision to SQL database access, using JDBC that has been returned queries from the database**.**

Matlab

Also known as Matrix laboratory is amongst the versatile and powerful data mining tool, that can be adopted to examine data though algorithm, model and application development. Also it can be applied as a stand-alone data mining tool rather than been in conjunction with other data mining tools.

Orange

It's an open source data mining and machine learning software that can be used virtualization and explorative data analysis. It provides a platform for recommender systems, predictive modelling and experiments analysis. Furthermore it can be applied in biomedicine, bioinformatics, teaching and genomic research. It is allow preferred to deal with new emerging innovations, the quality there of and the rollability of that emerging technology**.**

Tanagra

It's an open source software dedicated to researchers, students and academics for purely research purpose. The methods it proposes ranges from data analysis, machine learning, statistical learning and areas of databases research. Its objective is to afford researchers, academics and student the easiest way to use data mining tools to do analytics on synthetic or real data**.**

Knime

Also known as (Konstanz Information Miner) is an open source platform which focus mainly on data processing, integration, exploration and analysis. It uses the concept modular data pipelining in order to integrate different machine learning and data mining components.

4. METHODOLOGY AND APPROACH

The below objectives are set for this heart disease monitoring, prediction and detection system prototype and simulation.

- The prediction, detections and simulation systems shouldn't have assumptions about patient's records prior knowledge that it intends to compare.
- The system that is been selected for this test must be able to accommodate scalability against big databases that contains thousands of dataset.

The WEKA Tool

The approach choose for this research work, is been implemented using WEKA tool. WEKA can be regarded amongst the open source software tool that is tailored for data mining .and the accumulation of machine learning algorithms. Weka consists of techniques for pre-processing, clustering regression, classification and visualization. (4). The explore mode is normally used to test for classification. For this study a decision tree containing 10 folds will be used.

In order to run the test the following WEKA steps needs to be carried out.

- Start by clicking Weka explorer
- Open the saved dataset file in ARFF format.
- Click on the classify tab and select J8 etc. (from the tree)
- from the choose button
- select the desired testing mode option
- then click start button to start generating results

The process can also be repeated for Logistic model tree algorithm and Random forest Algorithm.

is normally used to test for classification. For this study a decision tree containing 10 folds will be used.

In order to run the test the following WEKA steps needs to be carried out.

- Start by clicking Weka explorer
- Open the saved dataset file in ARFF format.
- Click on the classify tab and select J8 etc. (from the tree)
- from the choose button
- select the desired testing mode option
- then click start button to start generating results

The process can also be repeated for Logistic model tree algorithm and Random forest Algorithm.

Data

For comparing various decision tree classification technique, the Cleveland dataset adopted from UCI repository is used, the dataset are available from at http://archieve.ics.uci.edu/ml/datasets/heart+disease. Originally the dataset consists of 76 attribute and 303 records. However for the purpose of this research sty only 13 attribute will be used for testing.

5. RESULTS

5.1 J48 with Reduction Error pruning Algorithm

The j48 algorithm sample is connected on to the Cleveland UCI repository, and the classification correctness and confusion matrix is generated from class having 2 conceivable qualities i.e. absent and present. From the confusion matrix a precise classification can be produced. The J48 algorithm results are projected in the figure below.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 111  39 |   a = absent
  29  91 |   b = present
```

Figure 1. Classification results for J48

	Train Error	Test error
J48	0.2636	0.2281

J48 model error rate is below 0.3 percent which is still acceptable

5.2 Logistic Model tree Algorithm

The Logistic model tree algorithm example is connected to the Cleveland UCI repository and the forecasted confusion matrix for two heart disease probability (.i.e. absent and present) are indicated in the figure below.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 131  19 |   a = absent
  22  98 |   b = present
```

Figure 2. Logistic Model tree classification results with confusion matrix

	Train Error	**Test error**
Logistic Model tree Algorithm	0.2149	0.2302

The train and test error are both below 0.3 percentage which is still an acceptable error range.

5.3 Radom Forest algorithm

```
=== Confusion Matrix ===

   a   b   <-- classified as
 150   0 |   a = absent
   0 120 |   b = present
```

Figure 3: Random forest Algorithm classification with confusion matrix.

	Train Error	Test error
Random Forest	0.0978	0.2696

6. COMPARISON OF METHODOLOGIES

	J48	Logistic model tree algorithm	Random Forest Algorithm
Train error	0.2636	0.2149	0.0978
Test error	0.2281	0.2302	0.2696

The comparison results on J48 algorithm, LMT and Random forest algorithm, Random forest achieve the highest accuracy with lower error on training. Therefore Random forest algorithm has the best overall performance.

CONCLUSION

When analysing the experiments results, it can be concluded that Random forest algorithm turn out to be the best decision tree classifier for predicting heart disease since it contains higher percentage of accuracy and a total of 0.04 second to build up. In conclusion , more multiple test using other classifier technique such as Naïve Bayes, K-nn and SVM needs to be done in order arrive at an informed decision in terms of the best classifier for predicting heart diseases using weka and other tools such as Matlab, orange etc

References

[1] G. Public and H. Security, “A safer,” 2007.

[2] C. Helma, E. Gottmann, and S. Kramer, “Knowledge discovery and data mining in toxicology,” *Stat. Methods Med. Res.*, vol. 9, no. 4, pp. 329–358, 2000.

[3] M. Lateef, Z. Iqbal, M. S. Akhtar, A. Jabbar, M. N. Khan, and A. H. Gilani,

"Preliminary screening of Trachyspermum ammi (L.) seed for anthelmintic activity in sheep," *Trop. Anim. Health Prod.*, vol. 38, no. 6, pp. 491–496, 2006.

[4] K. Manivannan, A. G. Rowan, Y. Tanaka, G. P. Taylor, and C. R. M. Bangham, "CADM1/TSLC1 Identifies HTLV-1-Infected Cells and Determines Their Susceptibility to CTL-Mediated Lysis," *PLoS Pathog.*, vol. 12, no. 4, 2016.

[5] A. A. Boxwala *et al.*, "Organization and representation of patient safety data: Current status and issues around generalizability and scalability," *J. Am. Med. Informatics Assoc.*, vol. 11, no. 6, pp. 468–478, 2004.

[6] C. S.Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, 2012.

[7] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, "Privacy-Preserving Patient-Centric Clinical Decision Support System on Naïve Bayesian Classification," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 2, pp. 655–668, 2016.

[8] A. Pattekari, S.A.; Parveen, "Prediction system for heart disease using Naïve Bayes," *Int. J. Adv. Comput. Math. Sci.*, vol. 3, no. 3, pp. 290–294, 2012.

[9] G. Komal and V. Vekariya, "An Approach to Secure Data in Disruption Tolerant Network," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 11, pp. 0396–0400, 2015.

[10] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 121, no. July 2014, pp. 23–30, 2010.

[11] K. U. Rani, "a Nalysis of H Eart D Iseases D Ataset Using," *Knowl. Manag.*, vol. 1, no. 5, pp. 1–8, 2011.

[12] D. B. U. R. Dhanalakshmi, "A Comprehensive Survey of Students Performance Using Various Data Mining Techniques," *Int. J. Sci. Res.*, vol. 6, no. 4, pp. 2233–2238, 2017.