Predicting Hepatitis Using Decision Trees

YAKUBANI YAKUBU[1], AUGUSTINE S. NSANG[2]

[1, 2] Department of Computer Science, School of Information Technology and Computing, American University of Nigeria, Yola By-Pass, Yola, Nigeria

Abstract- Hepatitis is a life-threatening disease caused by the swelling of the liver. It can be caused by an infection of the liver by hepatitis viruses (hepatitis B, hepatitis C, and viral hepatitis). Manual examination can be burdensome for large-scale diagnoses leading to severe economic impact to the individual health program. Automated hepatitis prediction using machine learning techniques such as decision tree, kNN and the perceptron offers the promise of serving an effective diagnostic aid. In this study a decision tree is constructed to predict hepatitis, and two other algorithms, kNN and the perceptron, are implemented to predict the same disease. The three learning algorithms are compared with each other by the extent to which they predict hepatitis. The comparisons are made using the RAND index and confusion matrices.

Indexed Terms- Data mining, Decision Tree, Hepatitis, KNN, Perceptron.

I. INTRODUCTION

Data mining is the process of extracting i.e. collecting, cleaning, processing, analyzing, and gaining useful information from data[2]. Data increases every day making it very difficult to derive useful information from the data. The process of extracting meaningful information from data is very important but it can be very tedious or challenging.

The extraction of a hidden pattern from data which is data mining is a beneficial method of analyzing data[15]. Data mining is not all about analyzing facts or patterns from raw data. It involves the establishment of practices and policies that manage the full data life cycle of an organization or enterprise. Data mining also involves the building of models and the deduction of inference [9].

Data mining techniques take their origin from machine learning, artificial intelligence, computer science, and statistics, etc[10].

Data mining involves several algorithms and techniques which include Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithms, Nearest Neighbor methods, etc. which are used for knowledge discovery in databases[5].

Machine learning is a branch of Artificial Intelligence (AI) which is based on the idea that the machine or system can learn, identify and also make decisions on data with little or no human intervention.

Machine learning uses several algorithms that repetitively learn from data to improve, describe data, and predict results from the data. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine learning model is an outcome derived from trained machine algorithms with data[11]. Machine learning algorithms are categorized as supervised and unsupervised learning algorithms that are used to resolve complex problems[14].

The task of the supervised learner is to predict the value of the function for any valid input after having seen a number of training examples (i.e. pair of input and target output). There are two main types of supervised learning algorithms: classification, and regression, where there are input and output, and the main role of the algorithm is to find a mapping between the input and the output. In classification, the task is to assign the training input to one of the predefined classes. In the simple case where we have two classes as in the spam email example, the predefined classes are (1 or 0) indicating the type of the email (spam or not spam), and the role of the algorithm is to classify the training examples to one of the two classes. The good learner is the one that can discriminate between the two classes perfectly if there are no different data points that have the same label and there are no identical points that have different labels[14].

In the unsupervised learning algorithm, the set of observations is categorized into groups or clusters basing the categorization on the similarity between them. This categorization is otherwise known as clustering. Many clustering algorithms exist, among which k-means clustering is the most famous for a large number of observations.

Classification is the process of defining a model that differentiates between class labels to provide the ability to use the model and predict the class of tuples whose class label is unidentified[20]. There are different classification techniques in data mining, which include Support Vector Machine (SVM), K-nearest neighbours, Naive Bayes, Artificial Neural Network (ANN), and Decision Trees. ID3 and C4.5 (also called J48) is a version of decision tree classification. The ease with which each of these techniques can be used to classify records depends on the nature of the pattern in data and the phenomena to investigate.

The perceptron is one of the oldest algorithms introduced by Frank Rosenblatt in 1957 which is used to classify each point of a dataset into either a positive or negative label[4]. It is an eager learner.

The K-nearest neighbours algorithm is one of the classification algorithms. It is a non-parametric method that has been used in the early 1970s in statistical applications. The basic theory behind kNN is that in the calibration dataset, it finds a group of k samples that are nearest to the unknown sample (e.g. based on distance functions). As a result, for this classifier, k plays an important role in the performance of the kNN, i.e. it is the key tuning parameter of kNN [18].

Hepatitis is a health predicament defined by swelling of the liver. It can be caused by an infection of the liver by hepatitis viruses (hepatitis B, hepatitis C, and viral hepatitis). It is a worldwide disease that caused the population a death rate of 1.34 million in 2015, more than HIV. It is estimated that each year 1.75 million people are infected with the hepatitis C virus (HCV). This disease leads to severe liver diseases in the world today including liver cancer[6]. Several methods have been used to perform analysis of the HCV life process to get the significant factors of the virus duplication process. The prediction of the virus can help biologists to design appropriate viral inhibitors [13].

Hepatitis may happen without any sign or symptoms which lead to the yellow colouring of the skin and increase of the spleen. The presence of jaundice in the body indicates the presence of liver disease. It may also lead to bodyweight loss. Some causes of hepatitis are bacterial, fungal and parasitic infections as well as decreased blood flow. Viral hepatitis is caused due to hepatotoxic viruses such as hepatitis A, hepatitis B, hepatitis C, hepatitis D, and hepatitis E. Alcoholic hepatitis is caused due to excessive intake of alcohol which leads to liver failure. Toxic and drug-induced hepatitis is caused due to the intake of chemical agents. Paracetamol is the leading cause of liver failure, which results in damaging the cell and structural changes. Autoimmune hepatitis is caused by an immune response [21]. Several tools have been used to diagnose the hepatitis disease but still have a deficiency of analyzing the biological data of Hepatitis illness in the world, which leads to the death of many[3].

The purpose of this research is to construct a decision tree that can be used to classify unclassified hepatitis patients.
The objectives of the research are to:

i. construct a decision tree model that can be used to predict hepatitis using some real datasets.
ii. implement the k-NN and perceptron algorithms.
iii. compare the 3 techniques by how accurately they can predict hepatitis using the RAND index and Confusion matrices.

As mentioned above, we will be working with real hepatitis datasets. The dataset we will use will be small in practice compared to those in hospitals. However, we are certain that our model will be able to work on datasets of all sizes.

The rest of the paper is organized as follows. In Section 2, we will describe the techniques used and explain how they are applied to achieve the objectives of the work. Section 3 presents the results obtained using the methodologies described in Section 2. We conclude the study in Section 4 as well as make recommendations for future work.

II. RELATED WORKS

In this section, we review related works, particularly those which involve predicting hepatitis using data mining techniques.

[6] compared and applied several machine learning algorithms to hepatitis C viral NS3 serine protease cleavage data. The study shows that machine learning algorithms can be used to determine the hepatitis C virus. The performance of these algorithms was tested using ROC-AUC curves to evaluate the performance of the models.

[13] predicted hepatitis C virus cleavage sites using a data mining approach known as a decision tree. The algorithm was implemented using classification and regression tree (CART) Mat lab toolbox with the GINI index as spitting criteria. Using the decision tree classifier model the prediction accuracy of 96% was achieved which was not the best, but the rules made by the decision tree prediction model made the achieved results more informative.

[21] presented a prediction system for the diagnosis of hepatitis using the decision tree algorithm. The algorithm used to construct the decision tree is C4.5 that concentrates on 19 attributes from the dataset which includes age, sex, steroids, antivirals, spleen, fatigue, malaise, anorexia, liver big, liver firm, spiders, bilirubin, varices, ascites, ALK phosphate, SGOT, albumin, time, and histology for the diagnosis of hepatitis. The accuracy of the algorithm is presented using the confusion matrix and the accuracy obtained from the result is 85.81%.

[3] researched investigating the hepatitis disease using different types of neural network algorithms such as Quick, Multiple, Dynamic and RBFN with different factors such as data size, learning cycle, and processing time to achieve the diagnosis accuracy and estimated error. The accuracy of the neural network was calculated using the confusion matrix.

[12] researched machine learning models to predict disease progression among veterans with hepatitis C virus. From the study, two machine learning was developed and compared to predict cirrhosis development in a large chronic hepatitis C infected data.

[1] conducted a study on the performance of machine learning approaches on the prediction of *esophageal varices* for Egyptian chronic hepatitis C patients. The study aimed to find solutions to diagnose the disease, by analyzing the facts gotten from data through classification analysis, using machine learning techniques for early prediction in cirrhotic patients based on their clinical examination. The model uses six machine learning algorithms such as Neural Networks, Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest and Bayesian Network to achieve their objective.

[19] presented a study on using machine learning algorithms to predict hepatitis B surface antigen *seroclearance*. Models were developed based on four algorithms, which include the extreme gradient boosting (XGBoost), Random forest (RF), decision tree (DCT), and logistic regression (LR). The best model was known by the use of the area under the receiver operating characteristic curve (AUC). The AUCs for the models use XGBoost with 89%, Random Forest with 82%, Decision Tree with 61%, and Logistics Regression with 68% respectively, with XGBoost giving the best predictive performance.

III. METHODOLOGY

A. Classification Techniques

Classification is one of the major techniques used in data mining for prediction of diseases. It is used to classify each disease stored as one row of a dataset into a predefined class such as “die” or “live”. Classes are sometimes called “labels” or “categories”. For this research, the following classification techniques will be used due to their flexibility and uniqueness.

- Decision Tree

A decision tree is a data classification technique that builds a classification model in the form of a tree-like structure based on hierarchy. It consists of nodes which are the root node, leaf nodes and branch nodes which represent objects based on their attributes [15].

The decision tree is a supervised learning algorithm. It is a representation of decision from a given set of attributes and it is deterministic.It is one of the most widely used classifiers in machine learning and statistics[14].

To construct a decision tree from a set of training data, we use the algorithm below. From this algorithm, we can see that to construct a decision tree we must carry out the following steps:

- Compute the entropy, H(D), of the database (i.e., training set)
- Compute the information content of the database (given as I(D) = 1 – H(D))
- Compute the Information Gain of each attribute using either the entropy or the information content

Algorithm 1:

```
GENERATEDECISIONTREE(Data, Node)
 $A_{max}$ = Attribute with maximum information gain
If G($A_{max}$)=0
   Then Node becomes leaf node with the most frequent
            class in Data
      Else assign the attribute $A_{max}$ to Node
 For each value a1,...,an of $A_{max}$,
        generate a successor node: K1,...,Kn
        Divide Data into D1,...,Dn with Di
              ={x∈Data|Amax(x)=ai}
        For all i∈{1,...,n}
              If all x∈Di belong to the same class Ci
              Then generate leaf node Ki of class Ci
              Else GENERATEDECISIONTREE (Di, Ki)
```

- The Entropy:

The entropy measures the uncertainty of a random variable; it characterizes the impurity of an arbitrary collection of examples. The higher the entropy the lower the information content.

If p is the probability distribution of an n-class dataset i.e. p = $(p1,\ p2,\ \dots,\ pn)$, then the entropy of this dataset can be computed as:

$$H(p) = H(p1,\ p2,\ \dots,\ pn) = -\sum_{i=1}^{n} p_i log_2 p_i$$

- Information Content:

The information content I(D) of a dataset D is the opposite of the uncertainty. As mentioned above, the higher the entropy the lower the information content.

Thus, information content of a dataset D is defined as:

$$I(D) := 1 - H(D)$$

where H(D) is the entropy of the Dataset.

- Information Gain:

The information gain $G(D, A)$ through the use of the attribute A is determined by the difference of the average information content of the dataset

$$D = D1 \cup D2 \cup \cdots \cup Dn$$

divided by the n-value attribute A and the information content $I(D)$ of the undivided dataset, which yields:

$$G(D,\ A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} I(Di) - I(D)$$

- Advantages of Decision Trees

i. Can generate understandable rules
ii. Perform classification without much computation
iii. Can handle continuous and categorical variables
iv. Provide a clear indication of which fields are most important for prediction or classification

- The k-Nearest Neighbors Algorithm (kNN)

The k-Nearest Neighbors (KNN) algorithm is a classification technique that uses ‘feature similarity’ to predict the values of new data points. This means that the new data point will be assigned a value based on how closely it matches the points in the training set[14].

The K-Nearest Neighbor algorithm is given below:

Algorithm 2: K-Nearest Neighbors

```
Input: D  (a training data set)
       D1 (a set of data points whose classes are to be predicted)
Output: D2 (a data set containing the test set and their corresponding predicted labels)
1. Compute the similarity between an observation in the test set and each observation in the training set.
2. Select the k nearest neighbours of the observation of the test set among the training samples
```

Assign to the test data the class which is most common among its k nearest neighbours

- The Perceptron

The perceptron is another supervised learning technique used in classifying a data set into either positive or negative. It was invented by Rosenblatt in the year 1958 [8]. It is a simple classification algorithm that can separate a dataset into two linearly separable sets of tuples. It is seen as a learning agent that maps a feature vector to a function value. The perceptron is made up of a summation processor that takes the dot product of the inputs and the weights and uses a one-step function to determine the output of the perceptron. Learning by the perceptron is completed when it happens that no error has occurred after an epoch (a complete pass through the training set) during the training phase. When the training is complete, the perceptron will respond, for any input presented to it, with an output that is the same as the output of the observation used in the training phase [7].

$$(f(x) = \begin{cases} -1\ if\ \ w.x < 0 \\ \ 1\ \ \ if\ \ \ w.x \geq\ 0 \end{cases}$$

(3.4)

The perceptron algorithm is given below:

Algorithm 3:

```
PERCEPTRON LEARNING [M+, M−]
w = arbitrary vector of real numbers
Repeat
    For all x ∈M+
        If w.x ≤ 0 Then w = w +x
    For all x ∈M−
        If w.x > 0 Then w = w − x
Until all x ∈ $M^+$ ∪ M− are correctly classified
```

B. *Measuring Classification Performance*

Confusion Matrix

A confusion matrix is a table used to describe the performance of a classification model or classifier on a set of test data for which the true values are known. It consists of information about the actual and predicted classification done by the classifier[17]. The table below shows a confusion matrix of a two-class classifier.

Table 1: Confusion Matrix of a two-class classifier

		Predicted	
		Negative	Positive
Actual	Negative	A	b
	Positive	C	d

The entries that make up the confusion matrix are as follows:

a. which is the number of negative observations correctly predicted as negative
b. which is the number of negative observations predicted as positive
c. which is the number of positive observations predicted as negative and
d. which is the number of positive observations correctly predicted as positive.

Some evaluation measures that can be obtained from a confusion matrix are:

- Accuracy (AC): This is the proportion of the number of correct predictions made by the classifier and is given by

$$AC = \frac{a+d}{a+b+c+d} \qquad (3.5)$$

- True Positive Rate (TP)/recall: This is the proportion of positive samples that are correctly classified and can be calculated as:

$$TP = \frac{d}{c+d} \qquad (3.6)$$

- True Negative Rate (TN): This is the proportion of negative samples correctly predicted as negative and can be calculated as:

$$TN = \frac{a}{a+b} \qquad (3.7)$$

- False Positive rate (FP): This is the proportion of negative samples incorrectly predicted as positive and can be calculated as:

$$FP = \frac{b}{a+b} \qquad (3.8)$$

- False Negative rate (FN): This is a proportion of positive samples incorrectly predicted as negative and can be calculated as:

$$FN = \frac{c}{c+d} \qquad (3.9)$$

- Precision (P): This is a proportion of the predicted positive instances that are correctly classified:

$$P = \frac{d}{b+d} \qquad (3.10)$$

RAND Index

The RAND index measures the extent to which the classification results of one technique agrees with those of another on the same dataset.

Suppose D is the original dataset and D1 and D2 are the results by classifying D using two different classification techniques.

For each u ∈ D1 we denote by f(u) the corresponding data point in D2

Define the following:
a: the number of pairs of rows (u, v) of D1 such that u and v belong to the same class in D1 and the corresponding rows f(u) and f(v) belong to the same class in D2

b: the number of pairs of rows (u, v) of D1 such that u and v belong to different classes in D1 and the corresponding rows f(u) and f(v) belong to different classes in D2

Define n_p = a + b. That is, n_p measures the total number of pairs for which the classification results of the two techniques agree. The rand index is defined using equation 3.3 below:

$$Rand = 100\frac{n_p}{|D|(|D|-1)/2} \qquad (3.3)$$

where |D| is the number of tuples of D1 or D2. The RAND index is the percentage of classification results in the original space and reduced space that agree, or, the extent (computed as a percentage) to which the classification of D is preserved in D1 [16].

C. Datasets

To achieve the aim and objective of this work online hepatitis datasets were used.

- Hepatitis Datasets

This hepatitis dataset used in this paper were obtained online from the “UCI machine learning repository”. The Dataset is available at https://archive.ics.uci.edu/ml/datasets/Hepatitis. The Hepatitis patients’ results based on the datasets are classified into two classes, either “Live” or “Die” based on predefined attributes. The hepatitis dataset contains 19 attributes and 155 data samples. For this work, due to the amount of time it takes to train the perceptron, 100 data samples are selected from the original hepatitis dataset with 14 attribute values. The classes are renamed as 1 and -1 (for Live and Die respectively) for use with the perceptron, and renamed as 1 and 2 (for Live and Die respectively) for use with the K-Nearest Neighbor classifier. Table 2 below shows the attributes of the datasets to be used in order of significance for this classification by the decision tree algorithm.

Table 2: Attributes from datasets used

S/N	Attributes	Output
	Age	10, 20, 30, 40, 50, 60, 70, 80
	Sex	Male, Female
	Steroid	TRUE, FALSE
	Antivirals	TRUE, FALSE
	Fatigue	TRUE, FALSE
	Malaise	TRUE, FALSE

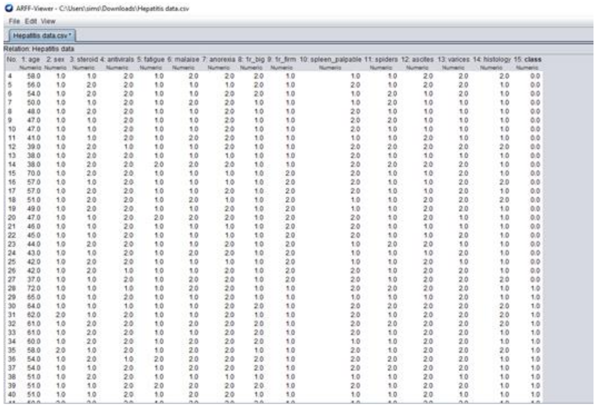
	Anorexia	TRUE, FALSE
	Liver big	TRUE, FALSE
	Liver Firm	TRUE, FALSE
	Spleen palpable	TRUE, FALSE
	Spiders	TRUE, FALSE
	Ascites	TRUE, FALSE
	Varices	TRUE, FALSE
	Histology	TRUE, FALSE
	Class	DIE. LIVE

IV. EXPERIMENTAL RESULTS

A. Decision Tree Analysis

The techniques discussed in the last section were implemented using the Weka and MATLAB platforms on the datasets mentioned above, and the results of analyzing the extent to which each of the classification algorithms can be used to predict hepatitis using the two methods of measuring classification performance will be presented in this section.

The decision tree obtained from the data is shown in Figure 3.2. This tree structure is built by considering 100 instances and 15 attributes from the datasets. This process is performed by using the Waikato Environment for Knowledge Analysis (WEKA) machine learning system to construct the decision tree. Figure 3.1 below shows the Datasets as viewed by the WEKA of the viewer.



Relation: Hepatitis data

No.	1: age	2: sex	3: steroid	4: antivirals	5: fatigue	6: malaise	7: anorexia	8: lr_big	9: lr_firm	10: spleen_palpable	11: spiders	12: ascites	13: varices	14: histology	15: class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
4	58.0	1.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	2.0	0.0
5	56.0	1.0	2.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	1.0	2.0	2.0	1.0	0.0
6	54.0	1.0	2.0	2.0	1.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	1.0	0.0
7	50.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	0.0
8	48.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	0.0
9	47.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	1.0	1.0	0.0
10	47.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	0.0
11	41.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	0.0
12	39.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	0.0
13	38.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	0.0
14	38.0	1.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	1.0	0.0
15	70.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	0.0
16	57.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	2.0	2.0	0.0
17	57.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	0.0
18	51.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	0.0
19	49.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	0.0
20	47.0	1.0	1.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	0.0
21	46.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	0.0
22	45.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	0.0
23	44.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0	0.0
24	43.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	0.0
25	42.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	0.0
26	42.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	0.0
27	37.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	0.0
28	72.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0
29	65.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0
30	64.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	1.0
31	62.0	2.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0
32	61.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	1.0
33	61.0	1.0	2.0	2.0	1.0	1.0	2.0	2.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0
34	60.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0
35	58.0	2.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0
36	54.0	1.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0
37	54.0	1.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0
38	51.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0
39	51.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0
40	51.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0

Figure 3. 1– Arff Viewer of datasets used in WEKA

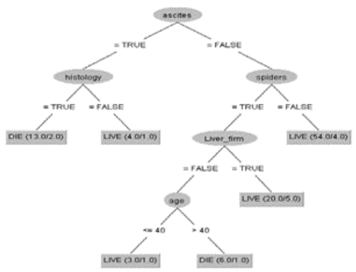


Figure 3. 2 Decision Tree

From the Decision Tree constructed above the confusion matrix for the decision tree algorithm is derived as shown below and from it, the accuracy, recall, precision, true positive rate and true negative rate are calculated.

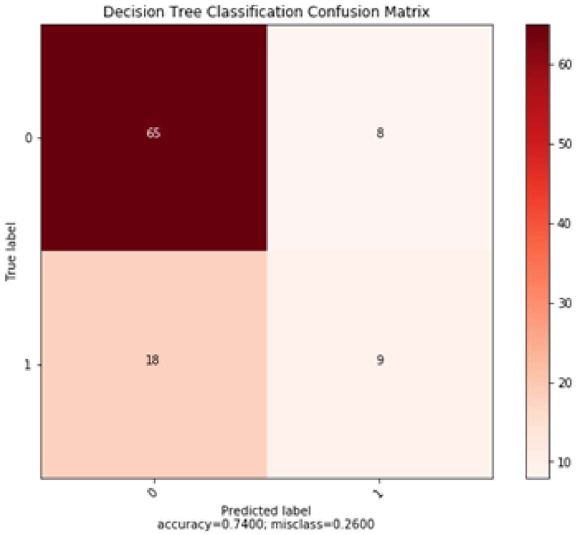


Figure 3. 3 Decision Tree Confusion Matrix

From the confusion matrix above it can be seen that the decision tree correctly classified 74% of the test samples and incorrectly classified 26% of the test samples.

Table 3 below shows the TP Rate, TN Rate, FP Rate, FN Rate and Precision based on the Confusion Matrix for the decision tree obtained using the data in Fig 3.1.

Table 3: TP Rate, TN Rate, FP Rate, FN Rate and Precision based on the Confusion Matrix in Fig 3.3

TP Rate	TN Rate	FP Rate	FN Rate	Precision
0.33	0.89	0.11	0.67	0.53

B. *k-NN Analysis*

Here, we present the results of the implementation of the K-Nearest Neighbors classification algorithm to predict hepatitis from an unclassified dataset. The classes are renamed to 1 and 2 (for Live and Die respectively) for use with the K-Nearest Neighbors classifier.

The Rand index is then used to evaluate the performance of the k-NN classifier. The results showed that the extent to which this classifier correctly predicts hepatitis for unclassified patients is 100%.

The confusion matrix is also used to evaluate the accuracy of the classification from the outcome of the implementation done using k-NN. The results obtained from the classifier is shown in the figure below:

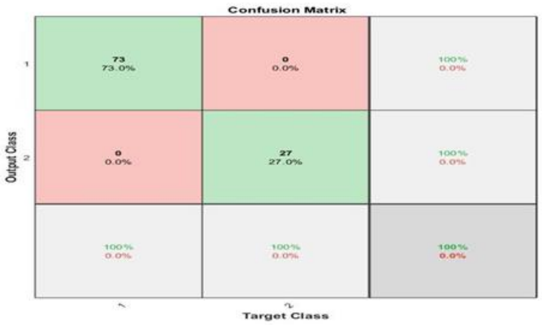


Figure 3. 4 KNN - Confusion Matrix

From the confusion matrix above it can also be seen that the k-NN classifier correctly classified 100% of the test samples.

Table 4 below shows the TP Rate, TN Rate, FP Rate, FN Rate and Precision based on the Confusion Matrix for the k-NN classifier obtained using the data in Fig 3.1.

Table 4: TP Rate, TN Rate, FP Rate, FN Rate and Precision based on the Confusion Matrix in Fig 3.4

TP Rate	TN Rate	FP Rate	FN Rate	Precision
1.00	1.00	0.00	0.00	1.00

C. *Perceptron Analysis*

In this section, the results of the implementation of the perceptron classification technique to predict hepatitis from the dataset are presented. The results are achieved by using the training set to obtain a weight vector, which is then used in classifying the unclassified dataset.

The Rand index is then used to evaluate the performance of the perceptron classifier. The results showed that the extent to which this classifier correctly predicts hepatitis for unclassified patients is 88.89%.

The confusion matrix is also used to evaluate the accuracy of the classification from the outcome of the implementation done using the perceptron classifier. The result obtained from the classifier is shown in the figure below:

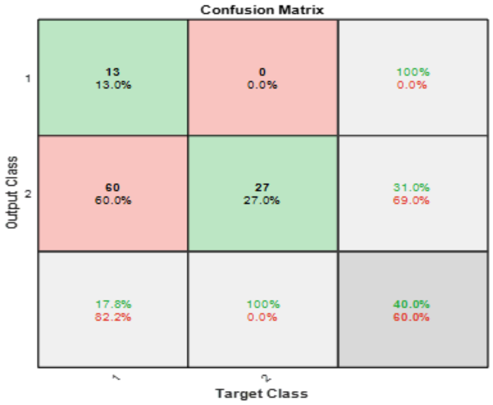


Figure 3. 5 Perceptron - Confusion Matrix

From the confusion matrix above it can be seen that the perceptron classifier correctly classified only 40% of the test samples.

Table 5 below shows the TP Rate, TN Rate, FP Rate, FN Rate and Precision based on the Confusion Matrix for the perceptron classifier obtained using the data in Fig 3.1.

Table 5: TP Rate, TN Rate, FP Rate, FN Rate and Precision based on the Confusion Matrix in Fig 3.5

TP Rate	TN Rate	FP Rate	FN Rate	Precision
0.31	1.00	0.00	0.69	1.00

Table 6: Comparing the three methods based on TP Rate, TN Rate, FP Rate, FN Rate, Precision and Rand index

	Metric Measures							
	Correctly Classified Instances %	Incorrectly Classified Instances %	TP Rate	TN Rate	FP Rate	FN Rate	Precision	Rand Index
Decision Tree	74.000	26.000	0.33	0.89	0.11	0.67	0.53	90.00
KNN	100.000	0.000	1.00	1.00	0.00	0.00	1.00	100.00
Perceptron	40.000	60.000	0.31	1.00	0.00	0.69	1.00	88.89

Rate, FP Rate, FN Rate, Precision and Rand index From the table above it can be seen in predicting hepatitis for an unclassified patient, the Decision Tree and KNN do better than the perceptron.

CONCLUSION AND FURTHER WORK

In this work, a Decision Tree to predict hepatitis was constructed. The tree was used to predict hepatitis for unclassified patients who will survive or die if they possess certain attributes of symptoms as indicated in the dataset. The K-NN and the perceptron algorithm were also implemented to predict this disease. As mentioned above, the confusion matrix and Rand Index were used to evaluate the performance of these algorithms and the results show that the Decision Tree and K-NN classifiers predict hepatitis better than the perceptron classifier.

We could speed up the classification algorithms using dimensionality reduction techniques. Also, ROC curves can be used to determine the extent to which each classification technique can correctly predict the classification of unseen data.

REFERENCES

[1] Abd El-Salam, S. M., Ezz, M. M., Hashem, S., Elakel, W., Salama, R., ElMakhzangy, H., & ElHefnawi, M. (2019). Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. Informatics in Medicine Unlocked, 17, 100267.

[2] Aggarwal, C. C. (2015). Data mining: The textbook. Springer.

[3] Alshamrani, B. S., & Osman, A. H. (2017). Investigation of Hepatitis Disease Diagnosis using Different Types of Neural Network Algorithms.

[4] Baba, U. A. (2017). Data Classification Using Various Learning Algorithms. MSc Thesis, 87.

[5] Bharati M., R. (2012). Data Mining Techniques and Applications (textbook), 301-305.

[6] Chown, H. (2019). A comparison of machine learning algorithms for the prediction of Hepatitis C NS3 protease cleavage sites. The EuroBiotech Journal, 3(4), 167–174.

[7] Ertel, W., & Black, N. (2011). Introduction to Artificial Intelligence. Springer.

[8] Haykin, S. S. (2009). Neural networks and learning machines (3. ed). Pearson.

[9] Jiawei, H., Micheline, K., & Jian Pei, P. (2011). Data Mining Concepts and Techniques.

[10] Jiechao, C. (2017). Data Mining Research in Education (textbook).

[11] Judith Hurwitz, D. K. (2018). Machine Learning for Dummies.

[12] Konerman, M. A., Beste, L. A., Van, T., Liu, B., Zhang, X., Zhu, J., Saini, S. D., Su, G. L., Nallamothu, B. K., Ioannou, G. N., & Waljee, A.

K. (2019). Machine learning models to predict disease progression among veterans with hepatitis C virus. PLOS ONE, 14(1), e0208141.

[13] Mohamed, A. (2011). A Data Mining Approach for the Prediction of Hepatitis C Virus Protease Cleavage Sites. International Journal of Advanced Computer Science and Applications.

[14] Mohamed, A. E. (2017). Comparative Study of Four Supervised Machine Learning Techniques for Classification.

[15] Mohanapriya, M., & Lekha, J. (2018). Comparative study between decision tree and knn of data mining classification technique. Journal of Physics: Conference Series.

[16] Nsang, A. Novel Approaches to Dimensionality Reduction and Applications: An Empirical Study. Lambert Academic Publishing, Saarbrücken, Germany, 2011.

[17] Santra, A. K., & Christy, C. J. (2012). Genetic Algorithms and Confusion Matrix for Document Clustering.

[18] Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery Sensors.

[19] Tian, X., Chong, Y., Huang, Y., Guo, P., Li, M., Zhang, W., Du, Z., Li, X., & Hao, Y. (2019). Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance. Computational and Mathematical Methods in Medicine, 2019, 1–7.

[20] Tribhuvan A. P, Tribhuvan P. P, & Gade J. G. (2015). Applying Naive Bayesian Classifier for Predicting Performance of a Student Using WEKA. 239–242.

[21] V.Shankar, S., V., S., C.P., K., & T.R., V. (2016). Diagnosis of Hepatitis using Decision tree algorithm. International Journal of Engineering and Technology.