6

Classification Model for The Survival of Patients with Hepatitis C Disease Using Decision Trees Algorithm

[1]Balogun, J.A., [1]Kasali, F.A., [2]Adekola, O.D., [3]Idowu, P.A., [1]Akinyemi, I.O.

[1]Department of Computer Science and Mathematics, Mountain Top University, Ibafo, Ogun State, Nigeria.

[2]Department of Software Engineering, Babcock University, Ilisan Remo, Ogun State, Nigeria.

[3]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.

Abstract:
Hepatitis C is an infectious disease caused by the hepatitis C virus and it primarily affects the liver. Hepatitis C often leads to liver disease and occasionally cirrhosis while in some cases, those with cirrhosis will develop complications such as liver failure. This study was aimed at developing a classification model for the classification of the survival of patients living with hepatitis C. The study collected data containing 19 attributes and 155 records from an online repository created by UCI machine learning repository following which C4.5 decision trees algorithm was adopted for developing the classification model. The results of the study showed that the decision tree is suitable for developing the classification model because of its structural representation of the classification of the survival of Hepatitis C disease. The evaluation of the model simulation process showed that the best performance using the percentage split was achieved using 90 percent for training and 10 percent for testing with an accuracy of 86.67% while using the k-fold cross validation was achieved using 10-fold with an accuracy of 83.87%. Overall, the best performance was achieved using 90 percent of the dataset for training and 10 percent for testing owing for an accuracy of 86.67%. The results also showed that a number of variables were extracted from the initially identified variables and were deemed more relevant for the classification of survival. The study concluded that the If-Then rules which were extracted from the decision tree proved effective in providing decision-support to experts due to its simplicity in interpretation thus mitigating deaths associated with hepatitis C disease.

Keywords: Hepatitis C, classification modeling, survival classification, decision trees algorithm.

1. Introduction

Hepatitis which is also referred to as the inflammation of the liver occurs in people without symptoms whereas in others develops a yellow discoloration of the skin and whitening of the eyes, poor appetite, vomiting, tiredness, abdominal pain, or diarrhea (Basra, Basra & Parupudi, 2011). There are five main types of viral hepatitis: type A, B, C, D, and E (Bernal & Wendon, 2013). Hepatitis A and E are mainly spread by contaminated food and water. Hepatitis B is mainly sexually transmitted, but may also be passed from mother to baby during pregnancy or childbirth (Dienstag, 2015). Both hepatitis B and hepatitis C are commonly spread through infected blood such as may occur during needle sharing by intravenous drug users. Hepatitis D can only infect people already infected with hepatitis B (Masouka & Chalasani, 2013).

Hepatitis C is an infectious disease caused by the hepatitis C virus (HCV) which primarily affects the liver (Centre for Disease Control CDC,

2016). During the initial infection people often have mild or no symptoms and occasionally a fever, dark urine, abdominal pain, and yellow tinged skin occurs. The virus persists in the liver in about 75% to 85% of those initially infected while earlier on chronic infection typically has no symptoms (Te & Jensen, 2010). According to the World Health Organization, WHO (2016), Hepatitis C often leads to liver disease and occasionally cirrhosis while in some cases, those with cirrhosis will develop complications such as liver failure, liver cancer, or dilated blood vessels in the esophagus and stomach. Hepatitis C is spread primarily by blood-to-blood contact associated with intravenous drug use, poorly sterilized medical equipment, needle-stick injuries in healthcare, and transfusions (Rutherford & Dienstag, 2016).

In the year 2015, the WHO estimated that 170 million people were infected with Hepatitis C globally and 3 to 4 million new infections occur each year, making it one of the leading public health problems in the world (Bernal & Wendon, 2015). With a prevalence of 5.3% and an estimated 32 million people infected with HCV, Sub Saharan Africa has the highest burden of the disease in the world. There is no vaccine against hepatitis C and its prevention includes harm reduction efforts among people who use intravenous drugs. Other prevention methods include: testing donated blood while chronic infection can be cured about 95% of the time with antiviral medications such as sofosbuvir or simeprevir. Hepatitis C infected individuals who develop cirrhosis or liver cancer may require a liver transplant and Hepatitis C is the leading reason for liver transplantation but, the virus usually recurs after transplantation (Basra, Basra & Parupudi, 2011).

Survival Analysis deals with the application of methods to estimate the likelihood of an event (death, survival, decay, child-birth etc.) occurring over a variable time period (Dimitologlou et al., 2012). Survival analysis is concerned with studying the time between entry into a study and a subsequent event (such as death). The traditional statistical methods applied in the area of survival analysis include the Kaplan-Meier (KM) estimator curve (Kaplan and Maier, 1958) and the Cox proportional hazard (PH) models (Cox, 1972). These methods apply parametric methods in estimating survival parameters for a group of individuals. Other methods applied in traditional statistical methods also include the use of non-parametric models. The Kaplan-Meier method allows for an estimation of the proportion of the population of people who survive a given length of time under some circumstances. The cox model is a statistical technique for exploring the relationship between

the survival of a patient and several explanatory variables.

Machine learning (ML) is a branch of artificial intelligence that allows computers to learn from past examples of data records (Quinlan, 1986). Unlike traditional explanatory statistical modeling techniques, machine learning does not rely on prior hypothesis (Waijee et al., 2013). Machine learning has found great importance in the area of classification modeling in medical research especially in the area of risk assessment, risk survival and risk recurrence (Cruz & Wishart, 2006). Machine learning techniques can be broadly classified into: supervised and unsupervised techniques; the earlier involves matching a set of input records to one out of two or more target classes while the latter is used to create clusters or attribute relationships from raw, unlabeled or unclassified datasets (Mitchell, 1997). Supervised machine learning algorithms can be used in the development of classification models. Classification modelling is aimed at allocating a set of input data records to a discrete target class.

Machine learning algorithms provide a means of obtaining objective unseen patterns from evidence-based information especially in the public health care sector using data mining (Idowu, Williams, Balogun & Oluwaranti, 2015). These techniques have allowed for not only substantial improvements to existing clinical decision support systems, but also a platform for improved patient-centered outcomes through the development of personalized prediction models tailored to a patient's medical history and current condition (Moudani, Shahin, Chakik & Rajab, 2011). To overcome this problem, medical decision support systems using data mining and machine learning is becoming more and more essential, which assists the doctors in taking correct decisions (Idowu, Balogun & Alaba, 2016). Feature selection methods are unsupervised machine learning techniques used to identify relevant attributes in a dataset. It is important in identifying irrelevant and redundant attributes that exist within a dataset which may increase computational complexity and time (Yildirim, 2015).

Among machine learning algorithms which are adopted for the development of classification models, the decision trees algorithm has gained more popularity because of its structural nature (Hall, 1999). Unlike most machine learning algorithms which are black-boxed models, the decision tree creates a top-down hierarchical structural tree which can be easily interpreted as a set of If-Then rules for interpreting the classification of a problem by a non-expert. In addition, the structural tree developed by the decision trees algorithm performs a feature selection of relevant variables which are adopted as the

modes of the tree thus providing information about the variables which have more importance than others for the classification problem. This study adopts the decision trees algorithm for the purpose of developing a classification model for the classification of the survival of patients with Hepatitis C and for the identification of the most relevant variables associated with survival, hence this study.

2. Related Works

Yasin, Jilani, and Danish (2011), worked on the development of a classification model for the classification of the diagnosis of hepatitis C disease using machine learning algorithms. The dataset of the study was collected from the University of California UCI Data Repository which consisted of 155 records consisting of 15 binary attributes, 5 continuous attributes and a class attribute. The results showed that the classification model developed using the dataset with relevant attributes performed better than the model developed using the initially identified attributes. The study was limited to the use of principal component analysis (PCA) for the identification of relevant attributes which is not accurate since biological information have a non-linear relationship unlike the linear relationship adopted by the PCA.

Agrawal et al., (2012), developed a classification model for the classification of the survival of the survival of lung cancer patients. Data for the study was collected from the Surveillance, Epidemiology and End Results (SEER) Program containing patients' data for survival of 6 months, 9 months, 1 year, 2 year and 5 years consisting of 13 input variables. Different decision trees algorithms were used for the formulation of the classification model, such as: C4.5 decision trees, random forest, Decision Stump and alternating decision trees. The decision trees algorithms used had accuracies of 73.61%, 74.45%, 76.80%, 85.45% and 91.35% for the 6 months, 9 months, 1 year, 2 year and 5 years survival dataset.

Idowu, Williams, Balogun and Oluwaranti (2015), worked on the classification of the risk of breast cancer using machine learning. The study collected primary data from a cancer registry in south-western Nigeria. The data collected consisted of information about non-modifiable and modifiable risk factors associated with the risk of breast cancer form healthy and infected women. The study adopted the use of the naïve Bayes classifier and C4.5 decision trees algorithm for the classification of the risk of breast cancer. The results of the study showed that the decision trees algorithm performed better than the naïve Bayes algorithm in the

classification of breast cancer risk. The study concluded that the variables identified by the decision trees algorithm improved the decision-making process of the medical experts by identifying a limited but important set of variables which improved the classification of the risk of breast cancer.

Balogun, Idowu and Oyekunle (2016), developed a classification model for the classification of the survival of Nigerian patients with Chronic Myeloid Leukaemia receiving Imatinib treatment. The study collected primary data from the medical records of a tertiary institution in south-western Nigeria consisting of 272 records. The study adopted the use of two decision trees algorithms namely: classification and regression trees (CART) and the C4.5 decision trees' algorithm for the classification of the survival of CML patients. The results of the study showed that the decision trees algorithm created a structured tree that was easily interpreted as a set of If-Then rules. The study concluded that the decision trees algorithms were able to extract the relevant variables that were most associated with survival classification.

Idowu, Balogun and Alaba (2016), worked on the prediction of the risk of infertility among Nigerian women using decision trees algorithms. The study collected information about the risk factors that were associated with infertility from infected and non-infected women. The study adopted the use of two decision tree algorithm namely: random trees and C4.5 decision trees algorithm. The results of the study showed that the C4.5 decision trees showed more performance compared to the random forest in the classification of infertility. The study concluded that the decision trees algorithm was effective in classification despite the use of small data records unlike other machine learning algorithms which required larger datasets for effective performance.

3. Materials and Methods

The dataset used in this study was collected from a public online repository called the UCI machine learning repository. The dataset consisted of information about the various variables which have been identified to be associated with the classification of the survival of hepatitis C patients receiving treatment alongside the target class which identified patients who survived (alive/Yes) and those who did not survive (dead/No). The data was collected from the URL https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/ which contained the dataset description and the dataset records. The hepatitis C datasets collected consisted of 19 input attributes and a target class called class and labelled as die for

those that did not survive and live for those that survived.

The dataset consisted of information that was collected from 155 patient records such that 32 patients did not survive (die) and 123 patients survived (live). The dataset collected was downloaded as two text (.txt) files from the online repository such that one contained the names and the other contained the data records. The two files were merged as one file and saved as a comma separated variable (.csv) file format following. Among the variables in the data collected, five of the variables are numeric while the remaining 13 variables were binary nominal variables.

The nominal variables include age (in years) which had three labels namely: below 30, between 30 and 50 and above 50; sex which had binary labels: male and female; while use of steroid, use of antivirals, presence of fatigue, malaise, anorexia, big liver, firm liver, palpable spleen, spiders, and ascites labelled as binary values yes and no including histology which was labelled as yes or no. The remaining variables, namely: level of bilirubin, alkaline phosphate level, sgot level (using sgot test), serum albumin level and protime level (using prothrombin time test) were all measured using numeric variables.

In this study, a classification model is required for the classification of the survival of patients with hepatitis C. The dataset collected consists of i input variables which are associated with the survival of the patients and contains information collected from j patients' records. Therefore, the dataset is a historical set of i attributes and j records defined as Xij which is mapped to the target class Yj. Therefore, the classification model which is required for the classification of survival is represented by a mapping, defined by equation (1) below.

$$\sigma: X \rightarrow Y \tag{1}$$

defined as: $\sigma(X_{ij}) = Y_j = \begin{cases} Live/Yes \\ Die/No \end{cases}$

3.1 C.45 decision trees algorithm

The theory of a decision tree has the following parts: a root node (represented by an input variable) was used as the starting point of the tree; then branches (edges representing label-values of variables) connect to other nodes below (Quinlan, 1986). Nodes that lie above other nodes are called parent nodes while those below are called child nodes. The terminal nodes (nodes with no child nodes) are called leaves (the target class). If-Then rules were induced from each complete top-down movement from the root node through branches that

connect to child nodes all the way to the terminal nodes (leaf). This path shows the flow of an If-Then statement following a top-bottom progressive pattern. Each node starting from the root node at the top are selected based on a heuristic metric that estimates how much information about the target class is possessed by the node thus its relevance.

The decision trees algorithm applies a divide-and-conquer technique for the process of the tree growth by adopting the hunt's algorithm. The basic idea of the decision tree analysis was to split the given dataset into subsets by recursive partitioning of the parent nodes into child nodes based on the homogeneity of within–node instances or separation of between-node instances with respect to their target class. Thus, at each node, the variables were examined using a metric and the splitter was chosen to be the variable such that after dividing the nodes into child nodes according to the value of the variable label, the target class is well differentiated. This process was repeated until no more division and creation of nodes is possible.

The decision trees algorithm chosen in this study required a heuristic criterion for optimal variable selection and split called the gain ratio (GR). Given a variable, X; the gain ratio (GR) relies on two equations, namely: the information gain (IG) and the split ratio. The information gain was used to measure the loss of entropy, H(X) of a variable. The entropy, H(X) of a variable, X is a measure of the number of bits required for storing information about the variable based on information available in the dataset. The entropy is determined as a function of the labels, $t \epsilon T$ of the variable, X_i according to equation (2)

$$Entropy, H(X) = -\sum_{t \epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|} \quad (2)$$

On the other hand, the information gain (IG) was used to determine the loss entropy as a result of the removal of the variable from the dataset. Thus, the higher the IG then the more the information loss then the relevant the variable is to be adopted as a node of the tree. The IG was determined by equation (3) and the value of the IG is normalized using a split ratio which in turn was used to determine the gain ratio (GR) by dividing the IG by the split ratio. The equation of the split

$$IG = H(X_i) - \sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \quad (3)$$

$$Split(X) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \quad (4)$$

3.2 Model simulation methods

Following the process of the identification of the decision trees algorithm considered in this study for the classification of the survival of patients with hepatitis C, there was a need to simulate and validate the classification model. This study considered the use of two training methods for the simulation of the classification model, namely: percentage split and k-fold cross validation methods. The percentage split required the dataset to be explicitly split into two parts such that a larger part was dedicated for training (building) the classification model while the smaller part was required for testing (validating) the classification model. The k-fold cross validation process required dividing the dataset into k-folds such that k-1 part was used for training while the remaining one part was used for testing. Therefore, 5 simulations were performed using each training technique such that for the percentage split 50%, 60%, 70%, 80% and 90% were used for training while the respective remaining were used for testing and 10-fold, 8-fold, 6-fold, 4-fold and 2-fold cross validation methods were adopted for the k-fold cross validation training techniques.

3.3 Model validation methods

In order to evaluate the performance of the supervised machine learning algorithms used for the classification of the survival of hepatitis C, there was the need to plot the results of the classification on a confusion matrix (Figure 1). A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. Correct classifications lie along the diagonal from the north-west corner to the south-east corner called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN) for storing the number of incorrect classifications made by the model. Therefore, number of actual yes/live cases are TP+FP, number of actual no/dead cases are FN+TN, number of predicted yes/live cases are TP+FP and number of predicted no/dead cases are FN+TN.

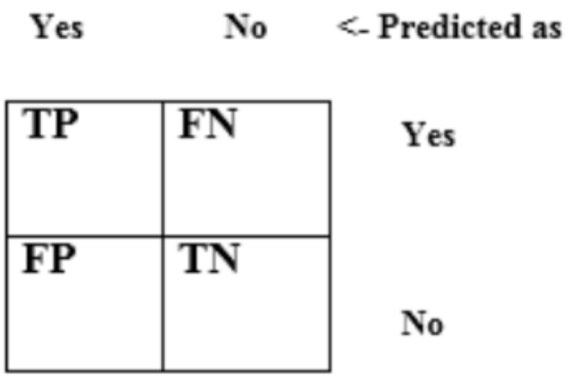

Fig. 1: Confusion Matrix for interpreting Model Simulation Results

The number of values within the four cells were in turn used to calculate four performance evaluation metrics that were used a basis of validating the performance of the decision trees algorithms that were developed from the ten simulations using the two different training methods chosen in this study. They are presented as follows:

a. Accuracy: was used to determine the proportion of records that were correctly classified by the decision trees algorithm expressed as a percentage. The closer this value is to 100 percent then the better.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (5)$$

b. True positive (TP) rate/sensitivity: was used to determine the proportion of the actual target class that was correctly classified by the decision trees algorithm. The closer this value was to 1 then

c.

$$TP\ rate_{Yes} = \frac{TP}{TP + FN} \quad (6a)$$

$$TP\ rate_{No} = \frac{TN}{FP + TN} \quad (6b)$$

d. False positive (FP) rate/false alarm: was used to determine the proportion of the actual target class that was incorrectly classified by the decision trees algorithm. The closer this value was to 0 then the better.

$$FP\ rate_{Yes} = \frac{FP}{FP + TN} \quad (7a)$$

$$FP\ rate_{No} = \frac{FN}{TP + FN} \quad (7b)$$

e. Precision: was used to determine the proportion of the predicted target class that was correctly classified by the decision trees algorithm. The closer this value was to 1 then the better

$$Precision_{Yes} = \frac{TP}{TP + FP} \quad (8a)$$

$$Precision_{No} = \frac{TN}{FN + TN} \quad (8b)$$

4. Results and Discussions

This section presents the results and discussions of the simulation and the validation of the classification model that was developed for the classification of the survival of patients with hepatitis C.

1.1 Results of percentage split simulation

The results of the model simulation process using the percentage split technique involved a process of model development by using a larger percentage of the dataset for training the model (training data) and a lower percentage for testing the model (testing data). By using a percentage of 90%, 80%, 70%, 60% and 50% for training a model, the results of the correct and incorrect classifications made by the C4.5 decision trees algorithm is presented in Figure 2 using 10%, 20%, 30%, 40% and 50% respectively of dataset for testing the developed model.

The results of using 90% for training the model is displayed on the confusion matrix shown in Figure 2 (top-left) which shows that 1 Die case and 14 Live cases were used for testing. The model developed using 90% of the dataset for training was able to correctly classify the 1 Die case and 12 Live cases but misclassified 2 Live cases as Die which also presented an accuracy of 86.7% for 13 correct out of all 15 actual cases. The results of using 80% for training the model is displayed on the confusion matrix shown in Figure 2 (top-center) which shows that 6 Die cases and 25 Live cases were used for testing. The model developed using 80% of the dataset for training was able to correctly classify 5 Die cases and 17 Live cases but misclassified 1 Die case as Live and 8 Live cases as Die which also presented an accuracy of 71% for 22 correct out of all 31 actual cases.

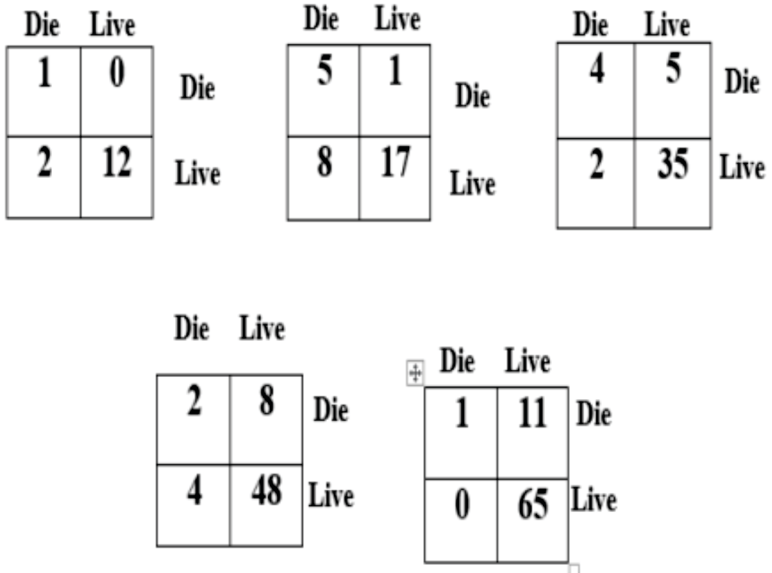

Fig. 2: Confusion Matrix for the Simulation based on Percentage Split Method

The results of using 70% for training the model is displayed on the confusion matrix shown in Figure 2 (top-right) which shows that 9 Die cases and 37 Live cases were used for testing. The model developed using 70% of the dataset for training was able to correctly classify 4 Die cases and 35 Live cases but misclassified 5 Die case as Live and 2 Live cases as Die which also presented an accuracy of 84.8% for 39 correct out of all 46 actual cases. The results of using 60% for training the model is displayed on the confusion matrix shown in Figure 2 (bottom-left) which shows that 10 Die cases and 52 Live cases were used for testing. The model developed using 60% of the dataset for training was able to correctly classify 2 Die cases and 48 Live cases but misclassified 8 Die case as Live and 4 Live cases as Die which also presented an accuracy of 80.6% for 50 correct out of all 62 actual cases.

The results of using 50% for training the model is displayed on the confusion matrix shown in Figure 2 (bottom-right) which shows that 12 Die cases and 65 Live cases were used for testing. The model developed using 50% of the dataset for training was able to correctly classify 1 Die case and all 65 Live cases but misclassified 11 Die cases as Live cases which also presented an accuracy of 85.7% for 66 correct out of all 77 actual cases.

4.2 Results of k-fold cross validation simulation

The results of the model simulation process using the k-fold cross validation technique involved a

process of model development by splitting the data into k folds following which k-1 folds are used for training while 1-fold is used for testing. The process of selecting 1-fold for testing was done by selecting 1-fold from the k folds with replacement from the first fold to the kth fold. By using 10, 8, 6, 4 and 2 folds for training a model, the results of the correct and incorrect classifications made by the C4.5 decision trees algorithm is presented in Figure 3 using 10, 8-, 6-, 4- and 2-folds cross validation respectively for the development of the model. Unlike the percentage split which uses a varying proportion of the total dataset for testing which gave rise to varying number of testing data, the cross validation uses a fixed number of testing data consisting of 155 records consisting of 32 Die cases and 123 Live cases

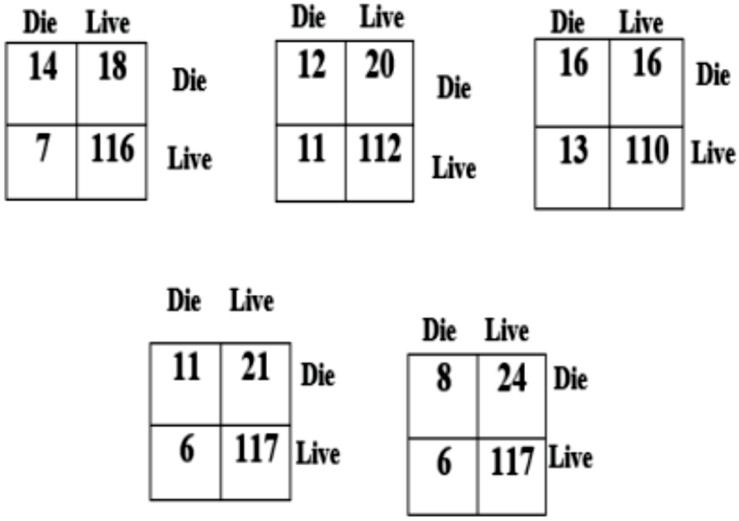

Fig. 3: Confusion Matrix for the Simulation based on k-fold Cross Validation

The results of using 10-fold cross validation for training the model is displayed on the confusion matrix shown in Figure 3 (top-left) which shows that 32 Die cases and 123 Live cases were used for testing. The model developed using 10-fold cross validation for training was able to correctly classify 14 Die cases and 116 Live cases but misclassified 18 Die cases as Live and 7 Live cases as Die which also presented an accuracy of 83.9% for 130 correct out of all 155 actual cases. The results of using 8-fold cross validation for training the model is displayed on the confusion matrix shown in Figure 3 (top-center) which shows that 32 Die cases and 123 Live cases were used for testing. The model developed using 8-fold cross validation for training was able to correctly classify 12 Die cases and 112

Live cases but misclassified 20 Die cases as Live and 11 Live cases as Die which also presented an accuracy of 80.0% for 124 correct out of all 155 actual cases.

The results of using 6-fold cross validation for training the model is displayed on the confusion matrix shown in Figure 3 (top-right) which shows that 32 Die cases and 123 Live cases were used for testing. The model developed using 6-fold cross validation for training was able to correctly classify 16 Die cases and 110 Live cases but misclassified 16 Die cases as Live and 13 Live cases as Die which also presented an accuracy of 81.3% for 1126 correct out of all 155 actual cases. The results of using 4-fold cross validation for training the model is displayed on the confusion matrix shown in Figure 3 (bottom-left) which shows that 32 Die cases and 123 Live cases were used for testing. The model developed using 4-fold cross validation for training was able to correctly classify 11 Die cases and 117 Live cases but misclassified 21 Die cases as Live and 6 Live cases as Die which also presented an accuracy of 82.6% for 128 correct out of all 155 actual cases.

The results of using 2-fold cross validation for training the model is displayed on the confusion matrix shown in Figure 3 (bottom-right) which shows that 32 Die cases and 123 Live cases were used for testing. The model developed using 2-fold cross validation for training was able to correctly classify 8 Die cases and 117 Live cases but misclassified 24 Die cases as Live and 6 Live cases as Die which also presented an accuracy of 80.6% for 125 correct out of all 155 actual cases.

4.3 Discussion of results

Based on the results presented earlier regarding the formulation and simulation of the results of this study, this section presents the discussion of the results presented. The results of the use of the percentage split and the k-fold cross validation technique for training the model provided a decision tree at the end of the simulation using C4.5 decision trees algorithm. The decision trees generated by the C4.5 decision trees algorithm is presented in Figure 4. The decision trees generated had as its nodes, attributes selected from the initially identified 19 attributes in the dataset collected for this study. Among the initially identified variables in this study, the variables used by the decision trees algorithm in generating the classification model for the survival of Hepatitis disease are: Presence of Ascites, Present Age of Patient, Presence of Spiders, Bilirubin content, Sex of the Patient, Firm Liver, Palpable Spleen, Big Liver and Presence of Anorexia.

These variables were used as the nodes (oval shape) generated by the decision trees algorithm that was used to build the tree which was composed

of 12 rules identified by the 12 leaf nodes (square shape) at the terminal point of the decision trees at the bottom. The rules were interpreted from the tree using If-Then rules to trace the relationship between the children's nodes from parent nodes all the way to the terminal nodes called the edges where the consequent part of is found.

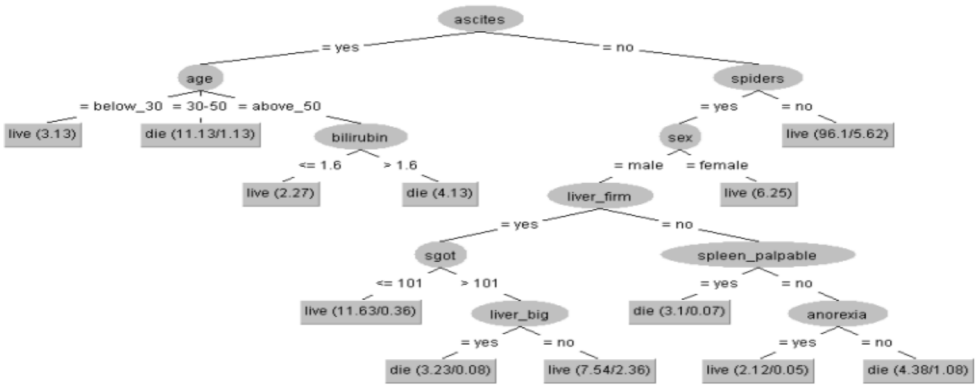

Fig. 2: Decision Trees for the Classification of Hepatitis C Survival

Following is a description of the rules extracted from the decision trees generate in this study. The extracted rules can be used to guide the clinical decision-making process taken by experts in the prognosis of the outcome of hepatitis patients

i. IF (Ascites=Yes) AND (Age=below 30) THEN (Survival=Live);
ii. IF (Ascites=Yes) AND (Age=30-50) THEN (Survival=Die);
iii. IF (Ascites=Yes) AND (Age=Above 50) AND (Bilirubin=<=1.6) THEN (Survival=Live);
iv. IF (Ascites=Yes) AND (Age=Above 50) AND (Bilirubin=>1.6) THEN (Survival=Die);
v. IF (Ascites=No) AND (Presence of Spiders=No) THEN (Survival=Live);
vi. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Female) THEN (Survival=Live);
vii. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=Yes) AND (Sgot=<=101) THEN (Survival=Live);

viii. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=Yes) AND (Sgot=>101) AND (Big Liver=Yes) THEN (Survival=Die);

ix. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=Yes) AND (Sgot=>101) AND (Big Liver=No) THEN (Survival=Live);

x. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=No) AND (Palpable Spleen=Yes) THEN (Survival=Die);

xi. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=No) AND (Palpable Spleen=No) AND (Presence of Anorexia=Yes) THEN (Survival=Live); and

xii. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=No) AND (Palpable Spleen=No) AND (Presence of Anorexia=No) THEN (Survival=Die).

Following the presentation of the decision tree that was generated in this study for the classification of the survival of patients with Hepatitis disease, the presentation of the discussion of the evaluation of the performance of the classification model based on the training techniques adopted in this study was presented. Table 1 shows a summary of the evaluation of the performance of the classification model developed using C4.5 decision trees algorithm via the percentage split and k-fold cross validation techniques. Using the percentage split, it was observed from the results that the best performance was achieved by using 90% of the total dataset records for training and using 10% for testing the performance of the model. It was observed that the model had an accuracy y of 86.7% as a result of 13 correct classification out of the 15 cases presented in the testing dataset. The results also showed that as the proportion of testing dataset was increasing from 10% to 50%, the accuracy of the model dropped to 71% using 80% for training, which increased to 84.8% using 70% for training, dropped to 80.7% using 60% for training and increased to 85.7% using 50% for training.

Table 1: Results of the Validation of the Decision Trees Model

Training Technique	Dataset	Correct/Total	Accuracy (%)	True Positive (TP) rate	False Positive (FP) rate	Precision	Area under ROC curve
Percentage Split Method	***90% Training***	*13/15*	*86.67*	*0.867*	*0.010*	*0.956*	*1.000*
	80% Training	22/31	70.97	0.710	0.196	0.836	0.777
	70% Training	39/46	84.78	0.848	0.457	0/834	0.652
	60% Training	50/62	80.65	0.806	0.683	0.773	0.786
	50% Training	66/77	85.71	0.857	0.774	0.878	0.712
K-Fold Cross Validation Technique	***10-Fold***	*130/155*	*83.87*	*0.839*	*0.458*	*0.825*	*0.757*
	8-Fold	124/155	80.00	0.800	0.514	0.781	0.603
	6-Fold	126/155	81.29	0.813	0.419	0.807	0.771
	4-Fold	128/155	82.58	0.826	0.531	0.806	0.630
	2-Fold	125/155	80.65	0.806	0.605	0.776	0.710

It was observed from the results that the best performance for the percentage split was achieved using 90% and 50% of the dataset records for training the model. However, using the 90% for training, it was observed that a better FP rate was achieved owing for an average of 1% of actual cases misclassified compared to using 50% for training. The results of the 90% for training also showed that an average of 87% of actual cases were correctly classified and an average of 96% of predicted cases were also correctly classified. Also, using the k-fold cross validation technique, it was observed from the results that the best performance was achieved by using 10-fold cross validation and using 4-fold cross validation. Using the 10-fold cross validation, it was observed that the model had an accuracy of 83.9% as a result of 130 correct classification out of the 155 cases presented in the dataset. The results also showed that as the number of folds reduced from 10 to 2, the accuracy of the model dropped to 80% using 8 folds, increased to 81.3% using 6 folds, increased to 82.6% using 4 folds and decreased to 80.7% using 2 folds.

It was observed from the results that the best performance for the k-fold cross validation technique was achieved using 10 folds and using 4 folds for training the model. However, using the 10 folds for training, it was observed that a better FP rate was achieved owing for an average of 45.8% of actual cases misclassified compared to using 4 folds. The results of the 10 folds also showed that an average of 83.9% of actual 63 cases were correctly classified and an average of 82.5% of predicted cases were also correctly classified. It was also observed from

the results that on a general note, the model with the best performance between those developed via percentage split and cross validation was the decision trees model developed using the 90% training dataset records. Unlike the model with the best performance among the k-fold cross validation technique which was the 10 folds which had an FP rate of 45.8%, the model developed using the 90% of dataset records for model training had an average of 1% of cases misclassified. Therefore, using the decision trees model for the classification of the survival of Hepatitis patients, clinical experts are able to make credible decision about patients.

5. Conclusion

This study identified variables that were related to the survival of Hepatitis patients receiving treatment and also collected relevant data from an online repository provided by the University of Chicago, Illinois (UCI) Machine Learning Repository. The study preprocessed the data collected from the repository for the purpose of formatting the dataset in order to be complaint with the tools proposed in this study. The study formulated a classification model for the survival of Hepatitis C patients using the C4.5 decision trees algorithm via a percentage split and k-fold cross validation techniques.

The study concluded that based on the variables that were identified in the dataset collected for this study, the C4.5 decision trees algorithm was formulated using a selected number of variables as the nodes of the generated tree. The study concluded that nine out of the initially identified 19 variables were relevant. The relevant variables selected in their order of importance for the classification of the survival of hepatitis C disease were: Presence of Ascites, Present Age of Patient, Presence of Spiders, Bilirubin content, Sex of the Patient, Firm Liver, Palpable Spleen, Big Liver and Presence of Anorexia. The set of relevant variables similar to those selected in the study by Yasin, Jilani and Danish (2011), were firm liver, big liver and anorexia.

In addition, 12 rules were extracted using IF-THEN rules based on the information about the none relevant variables extracted by the decision trees algorithm. The study also concluded from the results that using the 90% of dataset records for model building via the percentage split technique provided better classification results and lower misclassification results compared to other percentage split technique used and k-fold cross validation techniques. The study concluded that using a lesser number of variables for the classification of the survival of Hepatitis B patients on treatment will improve clinical decision making made by medical experts.

The study recommends that the classification model developed in this

study can be integrated into health information Systems in order to complement electronic health records systems which collect information about the identified variables and can be processed by the classification model for the identification of the clinical outcome of patients to whom treatment is provided.

References

Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L. & Choudhary, A. (2012). Lung Cancer Survival Prediction using Ensemble of Data Mining on SEER Data. *Scientific Programming, 20*: 29 – 42.

Balogun, J. A., Idowu, P. A. & Oyekunle, A. A. (2016): A Decision Trees-Based Classification Model for the Survival of Chronic Myeloid Leukaemia (CML) Patients. In Proceedings of the 10th International Conference on ICT Applications. Africa Centre of Excellence in Software Engineering, Obafemi Awolowo University, Ile-Ife, Osun State.

Basra, G., Basra, S. & Parupudi, S. (2011). Symptoms and Signs of Acute Alcoholic Hepatitis. *World Journal of Hepatology 3*(5): 118 – 120.

Bernal W. & Wendon J. (2013). Acute Liver Failure. *New England Journal of Medicine 369*(26): 2525 – 2534.

Centre for Disease Control, CDC (2016). Hepatitis C FAQs for Health Professionals. CDC. January 8, 2016. Retrieved on the 4th of February, 2018.

Cox, D. R. (1972). Regression models and Life Tables. *Journal of Stat. Soc. Serv. 34*: 187.

Cruz, J. A. and Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics 2*: 59 – 75

Dimitoglou, G., Adams, J. A. & Jim, C. M. (2012). Comparison of the C4.5 and a naïve bayes classifier for the prediction of lung cancer survivability. *Journal of Computing 4*(8): 1 – 12.

Hall, M. A. (1999). Correlation-based Feature Selection for Machine learning. PhD Thesis of the University of Waikato, Hamilton, New Zealand.

Idowu, P. A., Balogun, J. A. & Alaba, O. B. (2016): Data Mining Approach for Predicting the Likelihood of Infertility in Nigerian Women. Handbook of Research on Healthcare Administration and Management. Nilmini Wickramasinghe (Ed.) IGI Global Publishers. ISBN13: 9781522509202

Idowu, P. A., Williams, K. O., Balogun, J. A. & Oluwaranti, A. I. (2015): Breast Cancer Risk Prediction using Data Mining Classification Techniques. *Transactions on Networks and Communications, 3*(2), 1 – 11.

Kaplan, E. L. & Meier, P. (1958). Non-Parametric estimation from

incomplete observation. *Journal of American Statistical Association 53*: 457

Mitchell, T. (1997). *Machine Learning*. McGraw Hill, New York.

Moudani, W., Shahin, A., Chakik, F. & Rajab, D. (2011). Intelligent Predictive Osteoporosis System. *International Journal of Computer Applications, 32*(5): 28-37.

Quinlan, J. R. (1986). Introduction of Decision Trees. *Machine Learning, 1:* 81 – 106.

Rutherford, A. & Dienstag, J.L. (2016). Viral Hepatitis. In Greenberger, N.J., Blumberg, R.S. and Burakoff, R. *Current Diagnosis & Treatment: Gastroenterology, Hepatology, & Endoscopy*. 3rd Edition New York, NY: McGraw-Hill.

Te, H. S. and Jensen, D. M. (2010). Epidemiology of hepatitis B and C viruses: A Global Overview. *Journal of Clinical Liver Disease 14*(1): 1 – 21.

Waijee, A. K., Higgings, P. D. & Singal, A. G. (2013). A Primer on Classification models. *Clinical and Translational Gastroenterology 4*(44): 1 – 4.

World Health Organization, WHO (2016). Hepatitis C Fact sheet. Retrieved online on 4th February, 2018.

Yasin, H., Jilani., & Danish, M. (2011). Hepatitis C Classification using Data Mining Techniques. *International Journal of Computer Applications, 24*(3): 1-6.

Yildirim, P. (2015). Filter-Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. *International Journal of Machine Learning and Computing 5*(4): 258 – 263.