JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

Diagnosis of Hepatitis Disease Classification Using Non-Linear Compound Algorithms

Aqsa Jameel[1*], Imran Sarwar Bajwa[1], Mahvish Ponum[2], Tanzeela Kousar[3], Rabia Afzaal[4], Sajid Ali[5], Raheela Jameel[6], and Shahzad Jameel[7]

[1]Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.
[2]National University of Sciences and Technology, Islamabad, Pakistan.
[3]Institute of Computer Science and Information Technology, The Women University, Multan, Pakistan.
[4]Department of Computer Science, University of Lahore, Lahore, Pakistan.
[5]Department of Information Sciences, University of Education, Lahore, Multan Campus, Pakistan.
[6]Department of Chemistry, The Women University, Multan, Pakistan.
[7]Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan.
*Corresponding Author: Aqsa Jameel. Email: aqsa.6023@wum.edu.pk

Received: November 25, 2023 **Accepted:** January 31, 2024 **Published:** March 01, 2024

Abstract: The liver is a vibrant organ of our body that cast-off to process nutrients, scrimmage infections, and blood baptism. Inflammation of the liver causes an acute or chronic infection called hepatitis which happens when body tissues are contaminated due to the devouring of ethanol, some medication, and toxin. Detecting disease in its early stages can be difficult, as practitioners often struggle to predict the disease due to its ambiguous symptoms. The primary objective of this study is to develop a model for predicting liver disease in its early stages, which will aid practitioners in accurately diagnosing hepatitis. Different algorithms and techniques are available for data mining to solve data discovery problems and arrangement. The discussed algorithms are supervised learning whose labels are defined, in which classification is the vital method. However, this study focuses on the comparison of Classification and Regression tree (CART) and Java 48 (J48) using a 10-fold cross-validation method with comprehensive medical accuracy and well-informed decisions for disease detection. Amid these algorithms, decision trees are the simplest and easiest algorithms for understanding, decision making due to hierarchal structure in nature. The data set used in this analysis consisted of 155 patients with two classes and performance measures among said models. The comparison and investigation of the results revealed that the J48 algorithm shows improved performance with the highest classification rate and performance measures over CART obtained as an accuracy of 80%, a sensitivity of 88%, and a specificity of 52% to quantify how good and reliable the test is at detecting a positive disease. This article will aid physicians in classifying high-risk patients by making a novel prognosis, fending off and managing the disease by allowing data analysis of different patients by minimizing the need for excessive testing. Consequently, it will improve the patient's confidentiality by keeping them secure from health ramifications.

Keywords: Machine Learning; Decision Tree; Classification; Hepatitis Disease Detection.

1. Introduction

The liver is one of the most vital and solid organs in our body. The liver performs multiple functions such as it regulates blood clotting, maintaining sugar levels, and eliminating toxins from the body's blood supply.

It filters all of the blood and splits down poisonous substances such as drugs and alcohol. One of the important functions of the liver is to produce bile, which is fluid that helps to digest fats and keep away from waste. Swelling of the liver causes hepatitis [1]. Most of the people are symptom-free and even unaware of the infection. Other symptoms include fatigue, nausea, jaundice, loss of appetite, abdominal inflammation, and muddled thoughts. It is normally the result of a viral infection but there are other possible causes as well. Although its five main classifications are A, B, C, D, and E different virus is responsible for each type but frequently liver is influenced by hepatitis C which comes from Hepatitis C Virus (HCV) [2]. It causes liver damage priorly with the beginning of liver inflammation, gradually then leads to fibrosis or can say scarring as shown in Table 1. Fibrosis is a condition caused by our body's response to effect the liver so, it cannot be said that it is a disease. Shortly, it can be said that liver cells have been dead in hepatitis C. For optimum understanding, the stages of liver disease are discussed in Figure 1.

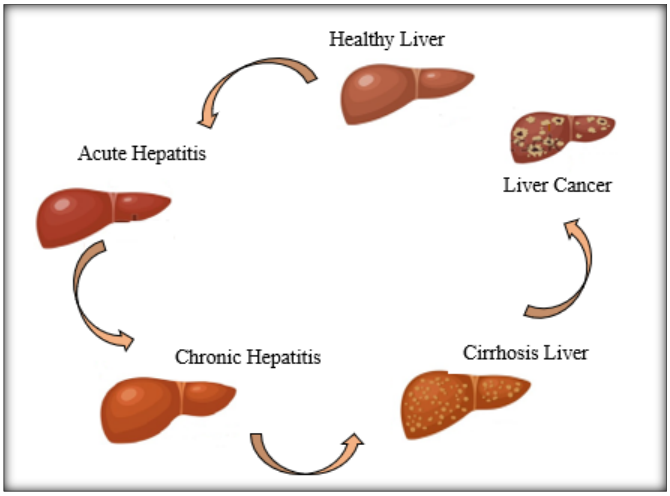

Figure 1. Stages of Liver Disease in Hepatitis

Machine Learning (ML) is the capability of computers to learn without being explicitly programmed and it is the study of algorithms that improve automatically with the aid of experiences. This domain focuses on the progression of computer programs when exposed to new data, can teach themselves to further grow and adapt to changes. It is close to computational statistics that focuses on making predictions after learning related the change.

Table 1. Stages of Chronic Hepatitis Disease with Fibrosis Condition

Stages	Description	Definition
0	No fibrosis	Generally, tissues are in normal connection
1	Portal fibrosis	Expand in fibrous portal
2	Periportal fibrosis	With minimum septa expansion into lobules/rare Porto-portal septa
3	Septal fibrosis	Architectural interference but no accurate cirrhosis
4	Cirrhosis	Diffuse nodular creation

Generally, machine learning refers to alteration in the system and the jobs related with artificial intelligence (AI). These jobs comprises on identification, analysis, planning, robotics control, forecasting, etc. "Change" may be for the production of a novel operating system or systems improvement [3].

Machine Learning is closely associated to statistics. Roslina et al. [4] suggested that ML discover and construct the study of algorithms that make prophecies on data. Input data is called training set while predictions and data-driven are considered as output. Machine Learning (ML) is used to analyze the computer to simulate human behavior study.

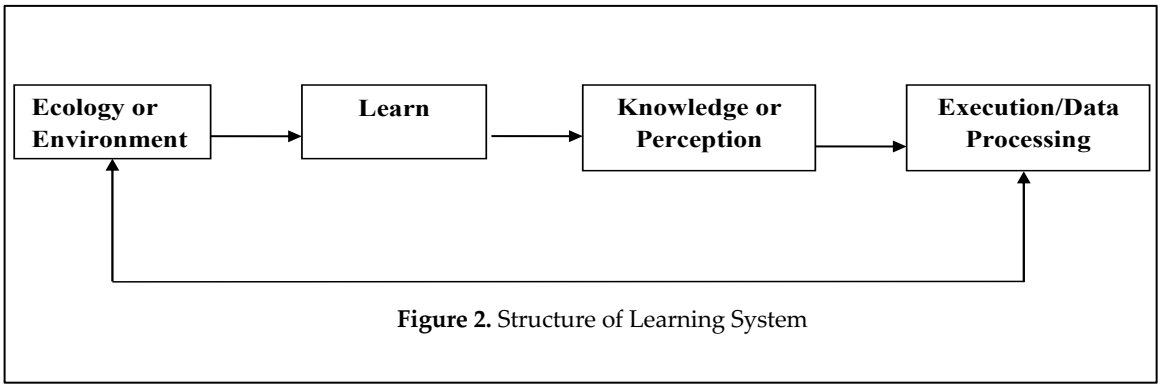

Figure 2. Structure of Learning System

The primary objective of the learning system as shown in Figure 2, is to boost the performance or the ability to get structured information that arranges structured data based on artificial intelligence. The structure of the learning system is a primary source of sense that makes the machine more intelligent [5].

Machine learning approaches are used for multiple fields that allows multiple systems to automatically generate the results by finding the pattern and relationship between the given data [6]. The medical industries collect a lot of data that is not mined properly and consequently there is not an ideal use. To search out such hidden patterns and the relationship between them often goes wrong. The study focuses on an intelligent clinical decision support system that will aid doctors in the medical field [7]. Intelligence can be expressed in many ways. "Ability of problem-solving is called intelligence, Capability to plan, think and scheduling is called intelligence, ability to cope up fuzzy and ambiguous problems, ability to understand, perceive, learn and recognize is called intelligence" [8]. AI approaches ranging from machine learning to deep learning are pervasive in the healthcare system for disease detection, patients risk identification, and drug discovery [9].

Machine Learning can be expressed in three categories.

Supervised Learning: Supervised learning comprises on training sets of examples. In this type, the input example is presented by a computer and outputs will be given by the teacher. The general principle is to map the input to give the output which is referred to as the destination. This kind of exemplary training set is offered by proper responses and all inputs are responded to by algorithms. [10]

This type of learning consists of classification & regression.

Unsupervised Learning: This is called feature learning or discovering hidden patterns in data because no labels are given to learning algorithms, to search structure in its input. Briefly, suitable responses are not offered in this type of learning.

In other words, it can be said that from the large stream of inputs, unsupervised learning is the ability to search the pattern from the given stream.

Reinforcement Learning: Reinforcement learning is between supervised and unsupervised learning. This type of learning does not propose perfection. An example of reinforcement learning is driving a car without a teacher. The algorithm gets notified when an answer is falsifies but it does not notify how to rectify this. This will work out how to search the valid answer.

If the answers are fair, the agent rewards good responses and is scold for bad ones.

Viral hepatitis disease has become a major public health concern on an all-inclusive scale. Chronic hepatitis can lead to high anguish and mortality. Early prognosis and treatment of disease can aid to minimize disease burden also with its transmission at the menace of infection or reinfection. In this context, this study has been conducted to propose a Clinical Decision Support (CDS) which is a twin of Machine Learning by using decision

trees, which will be a better solution. In the desertion of experienced doctors/physicians in rural, remote, and coastal areas, expert clinical decision support can assist and abet healthcare practitioners in examining patients' records and making trained and well-informed decisions for disease detection that can easily be filtered [11].

In addition, CDS can aid data regarding protocols of treatment by using Machine Learning algorithms. It also delivers updated alerts, promising patient care, and optimum decision-making, and employs knowledge management by offering clinical support in the form of suggestions. Thus, we aimed to develop a machine-learning model to predict hepatitis disease that could aid physicians in classifying high-risk patients and making a novel prognosis, fending off and managing the disease. As far as the objective of clinical support is concerned, it is purely responsible for assisting the physicians that allow data analysis of different patients that can be further used for designing the diagnosis. It is also vigilant the clinicians on time, which helps to minimize the costs while ameliorating the efficiency standards with appropriately reported errors [12].

The main contributions of this article are the following:

A detailed overview of Decision Trees supported by a clinical decision support system and its classification is presented. The data-preprocessing phase is compulsory to ensure that the instances of the data set are properly distributed in a balanced way that leads to effective classification results to predict the risk of hepatitis disease occurrence.

A data set feature analysis phase, consisting of three particular sub-steps: (1) Statistical description of attributes, (2) Important measurements by deploying decision tree classifiers, and (3) capturing variable characteristics or features a frequency of occurrence in tabular form from the data set. A comparative analysis of classifiers' performance is presented considering the most common parameters, such as sensitivity, specificity, True Positive, False Positive, Precision, Recall, F-Measure, time taken, corrected and incorrect instances, ROC Area, and Class. A performance evaluation is presented, where the J48 classifier achieves the higher results in all metrics, thus constituting the important suggestions of this analysis.

Hence, this article works only on the decision tree classification due to its hierarchical nature, diversity, and simplicity. Decision trees are also referred to as statistical classifiers due to their frequently used in classification issues by dividing the features into partitions, which help to minimize the recursion at every stage [13].

2. Literature Review

Multiple studies have been established to predict hepatitis disease detection by employing different classifiers and obtain the expected outcomes of their proposed system.

Bekir Karlik [14] presented comparative analysis between the two approaches. The first approach was by using naive Bayes and the second was backpropagation to detect hepatitis. These techniques compete very well with each other and also results are remarkable. The accuracy rate was 97% and 98% respectively. Open source software "Rapid Miner" was used for classification objectives. This software is used to analyze, estimation and forecasting of data [15].

Sathyadevi [16] worked on support vector machines & wrapper techniques. The patients affected by hepatitis are those who require particular and continuous medical treatment to minimize the mortality. By using Support Vector Machine (SVM) method, the prognosis of the hepatitis patient's life and its classification can be done. Well, the Wrapper Method is proposed to deminish extra features before moving toward the classification. This overall manuscript depicts that the Support Vector Machines can boost accuracy in aiding of selecting the features on a priority basis. In this study, the wrapper technique in WEKA for the prediction & classification of data execution using LibSVM.

Mahdieh Adeli et al. [17] aimed on medical prognosis via learning patterns with the help of collected data on hepatitis disease. The proposed method is used to establish such kind of system that will help physicians also gives a decision support system. In this paper, the use of CART, ID3, and C4.5 algorithms were proposed. Among these classifiers, binary decision trees were developed by utilizing the CART algorithm. It means that the decision tree which is generated by employing the CART algorithm has either two or no children as shown in Figure 3. Comparatively, the decision tree generated by the other one may have more children. For time

complexity and accuracy, the CART algorithm gives the efficient results as compared to other approaches like C4.5 and ID3.

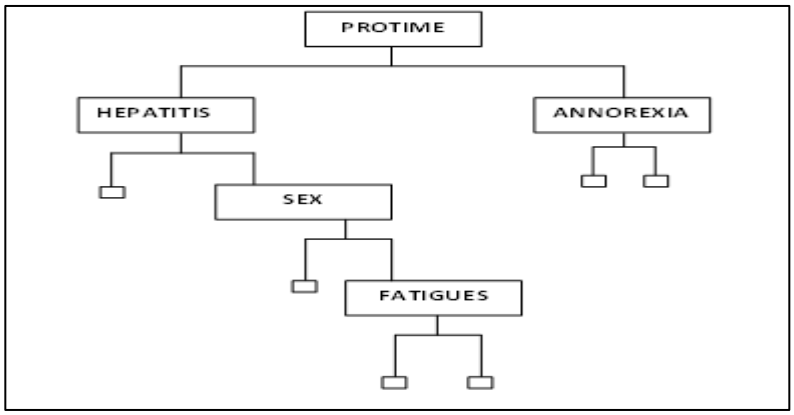

Figure 3. Decision Tree by CART Algorithm

Derya Serdar et al. [18] gives in depth understanding of an automated disease detection system for hepatitis disease on an "Extreme Learning Machine (EML)" by using pattern recognition as shown in Figure 4. In this manuscript, for the prognosis of hepatitis disease, an automatic intelligent system was presented. The proposed method can be appraised and judged by specificity analysis, sensitivity, and accuracy of classification. For the multiple type of activation function and the number of hidden or wrapped neurons, for the proposed method the accuracy was given by multiple equations. For the method mentioned, the best and well-suited classification accuracy was institute as 91.50%.

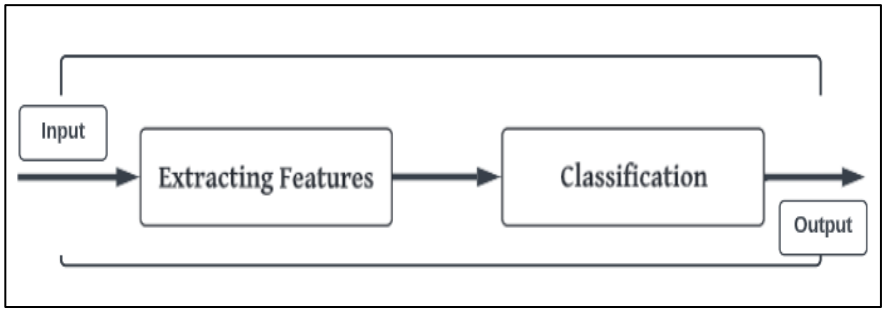

Figure 4. Pattern Recognition System

Fadl Mutaher Alwi et al. [19] applied classification and prediction in the discipline of bioinformatics. To analyze biomedical data, data mining techniques are widely used. Seven different algorithms named naive Bayes, naive Bayes updatable, FT tree, K-Star, J48, LMT, and neural network were reused to analyze hepatitis. WEKA is used as a data mining tool in this research. When the naive Bayes algorithm is applied to the hepatitis data set and compared to other algorithms, the time taken to run the data set is fast. So, naive Bayes sharpens the classification accuracy of the hepatitis dataset (Adel Nadjaran Toosi et al. [20].

Mohammed Afif et al. [21] focused to compare and analyze multiple techniques and their compatible tools as well as their impact on the healthcare industry. Transactions that take place in the healthcare industry produced a lot of problems and complexities. So, this paper also reduces the complexities of the transaction. A decision rule algorithm was used for data mining. The technique deployed as information gain and its accuracy level was 74%.

Sathyadevi [22] focused on "Single Nucleotide Polymorphism" data was identified by several machine learning techniques. To identify the SNPs related to the disease, this was integrated with "Feature Selection" algorithms as shown in Figure 5. Multiple techniques and algorithms were used in this paper like "Backtracking, forward selection, and backward elimination.

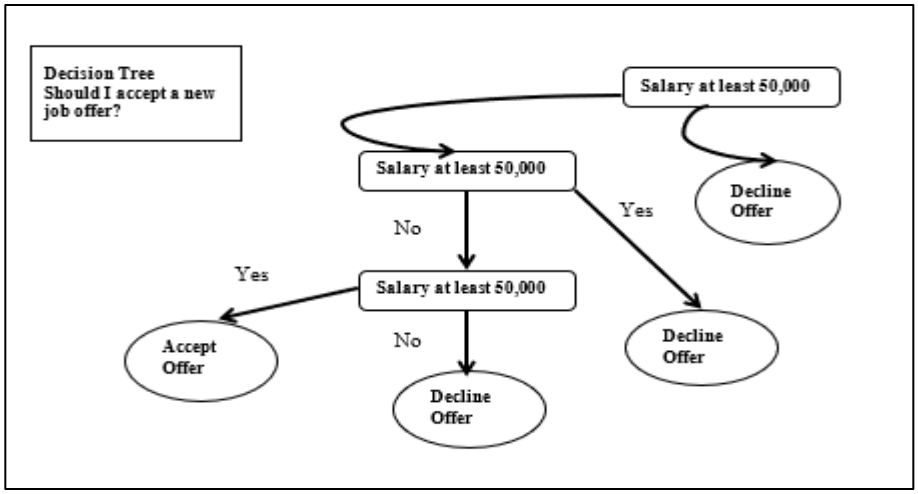

Figure 5. Decision Tree

Hui-Ling Chen et al. [23] established "Local Fisher Discriminant Analysis" and "Support Vector Machine" referred to as novel hybrid methods. Local Fisher Discriminant Analysis is used as a feature extraction tool to improve the diagnostic accuracy of the "Support Vector Machine" algorithm. The classification accuracy was 97% as compared to the other three existing methods PCA-SVM "Principle Component Analysis" SVM, FDA-SVM "Fisher Discriminant Analysis" SVM, and Standard SVM.

Okure Obot et al. [24] discussed a classification approach named "Scatter Search" abbreviated as SS combined with SVM "Support Vector Machine" for hepatitis disease diagnosis called 3SVM. The Scatter Search approach was used to search the optimal values of the support vector machine. Ten-fold cross-validation and holdout methods were used in this research. 3SVM gives best results than others where the average accuracy obtained from this is 99%. This paper includes 19 features. The purpose of the dataset is to forecast or predict the presence or absence of hepatitis disease.

Breiman et al. [25], present that traditional Chinese medicine has been vastly used to treat multiple diseases. In traditional Chinese medicine clinical practice "Treatment based on the Syndrome Differentiation" is the very basic principle. In nature, this is a decision-support problem that is based on a data warehouse. To cover different subjects of TCM, a series of a report called "Online Analytical Processing" as shown in Figure 6, on hepatitis and related system used to manage multiple reports related to hepatitis. OLAP reports are very beneficial for disease diagnosis and corresponding treatments for hepatitis.

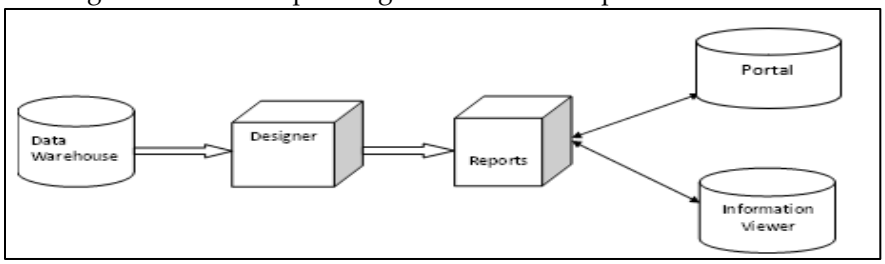

Figure 6. Realization Process of Online Analytical Processing (OLAP)

Carson K. Leung et al., [26] discussed the current globalized technological era; through which vast amounts of big data are collected and generated from a variety of data sources. These big data comes with variating levels in the sense that some data are brief while other are imprecise and ambiguous. Thus, knowledge discovered from these epidemiological data via data science approaches like machine learning, data mining and online analytical processing (OLAP) aids researchers as well as policymakers to have a better understanding of the disease which may motivate them to come up with ways to diagnose, control and combat the disease. Hence, it can be said that OLAP is sophisticated technology, which provides fast access to data analysis by using multidimensional hierarchical structures.

Pezhman Pasyar et al., [27] proposed a hybrid classification method (2021), for diffused ultrasound images of the liver by using Convolutional Neural Networks (CNN) for small data sets. In this study, transfer learning has used for the extraction of deep features with categorization for two class and three class classifiers for multiple networks, namely ResNext, ResNet 50, ResNet 18, Resnet 34, and AlexNet combined with fully connected networks are used. Experimental results obtain the highest accuracy for two class classifiers at 86.4% by using ResNet 50 with a hybrid classifier for liver disease. It can be concluded from this article to improve efficient and accurate results, larger data sets are necessary for fine-tuning.

Manjunath Varchagall1, D Sivakumar, et al., [28] developed a standardized model (2021), for the prognosis of early stages of liver disease with a Function Test (FT) to solve imbalance results by using two data sets from the UCI repository. For data set balancing, they used K Nearest neighbor (KNN) and Random Forest algorithms with the Synthetic Minority Oversampling method. For balanced and unbalanced data sets used in this manuscript, RF performs better over KNN in terms of false positive rate, the accuracy of the data set, specificity, and precision on balanced data sets whereas KNN gives efficient results over RF in terms of specificity, accuracy, sensitivity, false positive rate, and false negative rate parameters. Hence, it can be said that the stronger and more accurate results have driven for maximum parameters provided by a balanced data set. So, this approach helps physicians in correct and timely liver disease prediction in the early stages.

Further, in (2021) Niranjan Panigrahi et al., [29] suggested an experienced system refer to as a web-based Expert system shell which is deployed in an expert system builder consisting of 59 rules for Hepatitis-B. In this context, this web-based expert system is purely based on Clinical Decision Support System (CDSS). This research can assist health workers in rural as well as remote areas in the absence of skilled physicians to predict the disease. to check the efficiency and performance of this web-based system, testing has been carried out by questioning/querying the proposed system.

In (2022), Elias Dritsas and Maria Trigka [30] present a methodology, which is based on supervised learning to structure efficient models named probabilistic, tree-based, and ensemble learning for predicting kidney disease risk by evaluating support vector machine, linear regression artificial neural networks, and k nearest neural networks. The derived results obtained from SVM, LR, ANN, and k-NN give efficient performance as compared to other models with authentication of 100%, precision, recall, and f measure as 99.3%.

Sreenivasa Rao Veeranki and Manish Varshney (2022) [31] present a comparative performance analysis of chronic liver disease prediction with the help of a machine learning approach named Random Forest (RF), Multilayer Perceptron (MLP), K-Nearest Neighbour (K-NN) and Support Vector Machine (SVM) on Indian liver disease data set. After concluding the accuracy scores from all algorithms, it can be said that the support vector machine approach secures 74% accuracy compared to other classifiers. Hence, a suitable fitting approach is established in this manuscript. This work can be extended with other related techniques and possibilities that make a more accurate and efficient way for liver disease prediction.

Anusha Ampavathia and T. Vijaya Saradhi [32] designed a framework for multi-disease prognosis (2021), by using a hybrid deep learning approach for making accurate and feasible healthcare decisions. Multi diseases include hepatitis, diabetes, lung cancer, liver cancer, and heart disease. The proposed model consists of three phases namely data normalization, feature extraction, and prediction with the help of a meta-heuristic algorithm termed as Jaya Algorithm based Multi verse optimization algorithm for Deep Belief Networks and Recurrent Neural Networks. Hence, it can be said that after modification to a hybrid deep learning architecture weight can be optimized for both DBN and RNN by using the JA-MVO algorithm.

Jeong Hyun Lee et al., (2020) [33] proposed the automatic classification of liver fibrosis by using the context of a Deep Convolutional Neural Network (DCNN). They compared the performance classification of cirrhosis between the deep convolutional neural network and five radiologists. According to the results, the DCNN predicts maximum accuracy and also obtains more efficient performance than that of the radiologist in the detection of cirrhosis.

Neha Tanwar1 and Khandakar Faridar Rahman [34] present a very refreshing concept (2021). They proposed an automated decision-making system with its tools in the healthcare industry. They have used the

concepts of Big Data (BD), Machine Learning (ML), and Deep Learning (DL) by extracting information from huge data sets that will help physicians for accurate and timely decisions in making predictions and detecting diseases. This article also provides a remarkable review of the progression of Artificial Intelligence in detecting liver disease.

After reviewing this article, it is evident that Machine Learning techniques are highly promising with an optimized solution in detecting liver disease. The limitation of this article is that it needs further data to provide its validity, accuracy, and efficiency required for constant use by doctors and physicians.

Table 2. Comparative Analysis of Proposed Method with Deep Learning Methods

Ref.	Methodology	Characteristics	Challenges
Aliberti et al. (2019)	Feed Forward Network (FNN) and Recurrent Neural Network (RNN)	The networks can generalize the whole network by reading only a few data sets. It does not restrict/impose on input data.	FNNs require long training sessions. There are practical issues that may arise in Recurrent Neural Networks (RNN).
Almansour et al. (2019)	Support Vector Machine (SVM)	SVM has more accuracy than Artificial Neural Networks. It saves information on the entire network.	SVM is not well suited for large applications due to hardware dependency that affects the overall performance.
S. Mohan et al. (2019)	The Hybrid Random Forest With a Linear Model (HRFLM)	It can make enhanced prognosis with an effective accuracy level which improves the classification process (Li et al. 2017)	It is a time-consuming algorithm that leads to complexity. HRFLM is difficult to compute when compared to decision trees (Rao and Kumar 2013)
S. Hashem et al. (2018)	Multi-linear regression & Generative algorithm (GA) models, PSO, DT.	Easier data identification Accuracy could be maintained. Minimum computation time & robustness to overfitting.	The convergence rate is low. GA is very expensive in computation. Problem debugging is a quite difficult task.
Khan et al. (2019)	Classification of the tree with optimized parametric approach, Predictive regression.	Good prediction accuracy. It supports/promotes data equality.	It is not suitable for real-time data sets. Difficult to handle. Limited resources are not enough; more resources are required to tackle huge administrative data sets.
Baiying Lei et al. (2020)	Ensemble learning & Deep polynomial network (DPN)	Gives accurate predictions. Beneficial when data is not in a specific structure.	For large data sets, problem formulation is difficult (Hou et al. 2016). Optimization issues occur when the network becomes deeper.
Jiang et al. (Jiang et al. 2019)	Uses multi-task learning	Multitasking gives consistent results. It can prevent overfitting issues.	For large networks & experiments, this method is difficult to achieve. It could not work on multiple modalities of data.

Mohd Usama et al. (2019)	Recurrent Neural Network (Xue and Chuah 2018)	It can obtain high prediction accuracy. RNN can tackle any data size.	Computational complexity is maximum due to its recurrent.
El-Houssainy A. Rady et al. (2019)	Support Vector Machine (SVM) Probabilistic Neural Network (PNN) Radial Basis Function (RDF) Multilayer Perceptron (MLP)	PNN provides the highest classification accuracy as compared to other networks. Efficient prediction performance.	For new classification cases, PNN is slower than MLP networks. PNN requires maximum/more space to store the model.
Pezhman Pasyar et al., (2021)	Convolutional Neural Network by using the transfer-learning technique.	Provides hybrid approach Suitable for the small data set Highest accuracy for two-class classifier	Not suitable for the three-class classifier. The data set size should be larger to achieve maximum accuracy.
Niranjan Panigrahi et al., (2021)	Web-based Expert System	59 rules related to the knowledge base are used for designing.	Restricted rule base For more effective prediction, procedural knowledge can be maximized.

After analysis of Table 2, we can say that deep learning is an ongoing leading machine learning race consisting of better algorithms, computing power, and huge data. Still, ML attic and classical algorithms have a powerful position in the field [35].

Besides all this, there are plenty of ML classifiers too. No single ML algorithm works efficiently for all scenarios. Some of the parameters that affect our choice of taking into account an ML algorithm are the size of training data, accuracy, time taken, linearity, and no. of features. It can be said that a decision tree is typically optimum for small data sets, whereas deep learning algorithms often perform efficiently for large data sets [36].

ML is a knowledge base and scientific technique where computers learn how to sort out a problem, without programming them. CART and J48, both classifiers are simple decision trees non-parametric supervised machine learning techniques used for regression and classification issues. For that reason, every data scientist and ML engineer must have adequate knowledge of decision trees. [37]. After studying other approaches, decision trees take minimum effort for data preparation. However, to predict the target variable users need to have ready information to erect new variables. Decision trees can also create data classification without having to compute complicated calculations [38]. Inexplicably, you can visualize the DT. No preprocessing is needed as you do not need to get ready the data before building the model. These classifiers also provide data robustness as algorithms handle all types of data gently. It supports automatic feature interaction whereas another model like k-Nearest Neighbour (KNN) cannot. Therefore, the decision tree (DT) is faster as compared to KNN. [39]

The supremacy of decision tree classifiers in contrast with other machine learning methods is given below.

2.1 Decision Tree over Random Forest

A decision tree integrates some decisions, whereas a random forest integrates multiple decision trees. Hence, it is a lengthy process, yet slow. Comparatively, DT is fast and works easily on linear and non-linear data sets with simple training whereas the RF model requires rigorous training [40]. RF is highly complex in terms of training time as compared to CART and J48 decision trees because, for each DT prediction, a random forest has to generate output for given input data [41]. However, RF is less prone to overfitting as compared to DT.

2.2 Decision tree over KNN

DT and KNN, both are non-parametric approaches. DT supports a facility for automatic feature interaction whereas KNN cannot. DT is also fasters than KNN due to costly real-time execution [42].

2.3 Decision Trees over Naïve Bayes

DT is a deterministic model while NB is a generative model. DT is quite simple and easy to use. Although naive Bayes outperforms the classifier more than the decision tree in terms of recall, f-measure, accuracy, and AUC however, a decision tree can better handle precision [43].

2.4 Decision Tree over Neural Networks

CART, J48 decision tree classifiers, and Neural Networks both search non-linear solutions and have reciprocity between independent variables. DT works efficiently where there is a large set of categorical values in data training whereas NN gives efficient results when there is sufficient training data. DT is also better than neural networks when the situation needs a detailed description of the decision [44].

2.5 Decision Tree over SVM

DT uses hyper-rectangles in input space to sort out the problem whereas SVM uses the kernel tricky method for solving non-linear problems. DT is suitable for categorical data as it handles collinearity up to the mark more than SVM. Additionally, another vigorous crux concerning classifiers is [45]: CART and J48 DT provide the optimum solution in capturing non-linear associations, which can be complex to achieve with other classifiers like SVM and Linear Regression. Provide convenience to people: It is a remarkable aspect of DT. Outputs generated are easier to read without demanding statistical knowledge of difficult concepts.Some people think DT is more closely a reflection of human decision-making power than other techniques like classification and regression.

CART and J48 DT can be visualized graphically and can be easily handled by non-technical or non-experts. Without creating dummy variables, both classifiers can handle qualitative parameters conveniently. DT is specifically designed for non-linear data sets. For CART and J48, no hectic preprocessing and assumptions are needed on data with understandable descriptions of the prediction. It takes minimum time and supervision for data preparation. It can also create data classification without having to compute complex calculations. Both classifiers proved excellent for data mining tasks due to minimum data pre-processing and training.

CART and J48, both are easier to interpret compared to other ML models as they are close to black boxes. Having said that decision trees also come with some limitations [45]. DT can be non-robust as a minor change in the data sometimes can cause a huge change in a final estimated tree while training complex data sets. As a tree maximizes in size, it becomes susceptible to overfitting. They cannot perform well on large data sets. Due to high variance in the model and thinner splitting by one variable, CART predictions may be unstable whereas J48 provides stable outcomes due to recursively multiple variable splitting criteria. To tackle this, other tree-based algorithms such as random forest and other boosting classifiers can be used but they lost interpretability.

Finally, this study gives us a comparative analysis of decision tree classifiers which are Java 48 (J48), Classification, and Regression Tree (CART) for accurate diagnosis of hepatitis disease detection. Having said that, this article is conspicuous, as no single previous analysis has been conducted for hepatitis disease detection by using these classifiers. Our study disseminate more efficient results than most of the previous studies, i.e., we accessed 80% accuracy within 0.06 s by using J48 whereas 75% accuracy by using CART within 0.19 s.

3. Materials and Methods

The decision tree family consists of multiple algorithms but we have chosen two algorithms, which are J48 and CART that are rarely used on this dataset and have diversity in their nature. Both algorithms are easy to use and belong to the same hieratical in nature by constructing a tree, which is a decision rule. The decision tree split the features into divisions, which helps to minimize the recursion or repetition at every stage belonging to the same stage. To analyze/examine the tree, the parent node is considered and it is used to predict the data of the new label. For better understanding, the pseudo-code of both classifiers is discussed in Figure 7 a,b. The working of the J48 model is as follows: The decision tree is constructed based on the values of attributes

in data set training. The attributes that have the peak information gain are selected. A child of the root is created for each possible value by ordering training instances to the appropriate child node. After that, the branch is terminated with a target value. Finally, the decision tree is created and the target value of the new instance can be predicted. Figure 7b illustrates a pseudo code of the J48 classifier. Whereas each fork is split into predictor variables and each node has a prediction for the target variable at the end as in Figure 7a.

(a) Algorithm 1: Pseudo code of the CART algorithm

```
e = 0, endtree = 0
Note (0) = 1, Node (1) = 0, Node (2) = 0
While endtree < 1
    If Node (2^e -1) + Node (2^e) + ……+ Node (2^(e+1) - 2) = 2-2^(e+1)
        endtree = 1
    Else
        do i = 2^e – 1, 2^e,……2^(e+1) -2
            If Node (i) > -1
                Split tree
                else
                Node (2i + 1) = -1
                Node (2i + 2) = -1
                end if
                end do
            end if
        e = e+1
    end while
```

(b) Algorithm 2: Pseudo code of the J48 algorithm

```
Input: a dataset I
start
Tree = {}
If (I is "pure") || (other stopping criteria met) then terminate;
For all attributes a belongs to I do
Compute criteria of impurity function if we split on a;
a best = Best attribute according to the above-calculated criteria
Tree = create decision node that tests a best in the root
Iv = induced sub-datasets from I based on a best
For all Iv do
Start
Tree v = J48 (Iv)
Attach Tree v to the corresponding branch of the tree
end
Return tree
end
```

Figure 7. Pseudo code of (a) CART, (b) J48 Algorithm

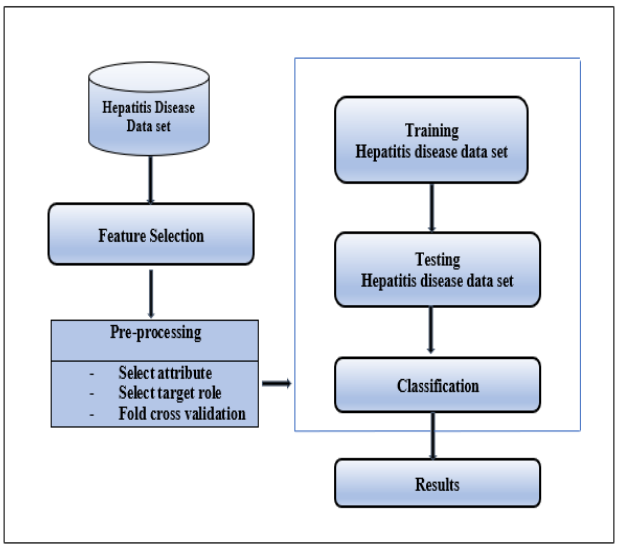

Figure 8. Pre-Processing Model

In this research, we have used Waikato Environment for Knowledge Analysis (WEKA) tool to analyze the classification of algorithms. The detailed work of WEKA is discussed below in Figure 9. This also involves coming out with a suitable model between J48 and CART that can apply to multivariate data sets providing efficient and accurate results and taking minimum time to build a model as shown in Figure 8.

3.1 Data Pre-Processing

Data preprocessing: it includes 1) Data cleaning, and transformation 2) Best variable selection for use in model building 3) Suitable classification for maximum prediction 4) Cross-validation for training and testing. For this, a total of 155 patients were included in this study of whom 32 patients are related to the first class (Die) and 123 patients belong to the second class (Alive). As preprocessing is a fundamental and necessary step in machine learning; therefore, we delete all those attributes that contained null values. In addition, to get a novel and high-quality statistical analysis of the data set, normalization, and data imputation were performed by using Chi-square Automatic Interaction Detector (CHAID) method that is used to discover the interactions between variables for tree classification [46]. In CHAID analysis, we can visualize the relationships that help to determine how variables best combine to define the outcome in a given dependent variable and generate synthesis samples for both classes. WEKA (version 3.8) tool was used to construct the model because it is an integration of visual tools and a graphical user interface for easily performing algorithms.

This section also discusses the further methodology steps taken in comparing J48 & CART.

- To analyze the comparison between CART and J48, the first step is to select the algorithm.
- Choose the data set with several instances are 155 for testing.
- Split the data set into 10 subsets for cross-validation (Mode of test: 10-fold cross-validation).
- Run cross-validation.
- Execute the test multiple times by a variating number of instances.
- Finally, summarizes the test results.

The numbers of attributes are 20 including the class attribute. The number of Instances is 155 as shown in Table 3. All the Missing Attribute values are denoted by "?".

Table 3. Features of Data Set

Features	Value
Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer, Real
Associated Tasks:	Classification
Number of Instances:	155
No of Attributes:	19
Missing Values:	Yes
Area:	Life

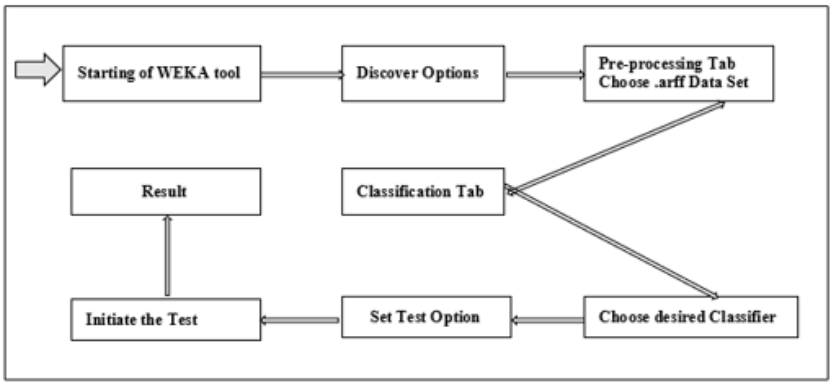

Figure 9. WEKA Data Flow Diagram

3.2 Data Understanding

3.2.1 Data Source

The data set is taken from the University of California Irvine (UCI) machine learning repository.

3.2.2 Data Preparation

The decision for strong prediction of hepatitis disease is based on various attributes and clinical factors that are Age, Sex, Steroid, Antiviral, Fatigue, Malaise, Anorexia, Liver big, Liver firm, Spleen palpable, Spider, Ascites, Varices, Bilirubin, Alk phosphate, Sgot, Albumin, Protime, Histology, and their class as shown in Table 4. There are two classes in Hepatitis Disease Dataset. The first class is Die which has 32 instances and the second class is Alive which comprises 123 instances in Table 5.

Table 4. Hepatitis Variables with Possible Values Used in Analysis

Attribute	Attribute Type	Class	Possible Value
Age	(Numeric)	Predictor	10 - 80 (Years)
Sex	(Nominal)	Predictor	Male, Female
std (Steroid)	(Nominal)	Predictor	Yes (1), No (0)
av (Antiviral)	(Nominal)	Predictor	Yes (1), No (0)
fg (Fatigue)	(Nominal)	Predictor	Yes (1), No (0)
mls (Malaise)	(Nominal)	Predictor	Yes (1), No (0)
ar (Anorexia)	(Nominal)	Predictor	Yes (1), No (0)
lb (Liver big)	(Nominal)	Predictor	Yes (1), No (0)
lf (Liver firm)	(Nominal)	Predictor	Yes (1), No (0)
sp (Spleen palpa-	(Nominal)	Predictor	Yes (1), No (0)
spd (Spiders)	(Nominal)	Predictor	Yes (1), No (0)
asc (Ascites)	(Nominal)	Predictor	Yes (1), No (0)
var (Varices)	(Nominal)	Predictor	Yes (1), No (0)
br (Bilirubin)	(Nominal)	Predictor	0.39, 0.80, 1.20,
ap (Alk phos-	(Nominal)	Predictor	33, 80, 120, 160, 200,
sg (Sgot)	(Nominal)	Predictor	13, 100, 200, 300,400,
al (Albumin)	(Nominal)	Predictor	2.1, 3.0, 3.8, 4.5,
pro (Protime)	(Nominal)	Predictor	10, 20, 30, 40, 50,
hist (Histology)	(Nominal)	Predictor	Yes (1), No (0)
Class	(Nominal)	Target	Die, Alive

Table 5. Class Distribution

Class	Number of Instances
Die	32
Alive	123

3.3 Classification Method

To classify new patients, our predicted model is based on the following indicators.

3.3.1 Binominal Classification

This type of classification consists of either predicted yes or predicted no values because the class attribute consists of two classes of problems. So, the algorithm predicts one of these i.e. a match may be detected/predicted or not.

3.3.2 Multinominal Classification

This type of classification is used where there are multiple classes/labels, i.e., [0 to N-1]. For such kinds of problems, the classifier predicts one of all these. In this study, CART and J48 classifiers are used to classify hepatitis disease. Both algorithms with their working are discussed in the following section.

3.4 Classification and Regression Tree (CART)

This is a decision tree algorithm that is used to detect disease, solve operational problems, and maximize efficiency and accuracy. Determination and selection of related attributes are challenging tasks. CART provides a method to handle missing values in a sophisticated way. The authors of the study, Fikes et al. [47] suggested that CART is an analysis tool for optimal handling of missing values of our incomplete data.

CART is used to solve the following type of problems:

1. Solve to maximize the tree problem, in short, it can be said how big to grow the tree.
2. Two ways splitting (binary form) strictly follow the rule. It means that data is split into partitions. Hence, partitions can also be split into further partitions or subnets. CART tree is generated by repeating partitioning of the data set.
3. CART algorithm can handle missing values.
4. Validate the tree and test automatically.

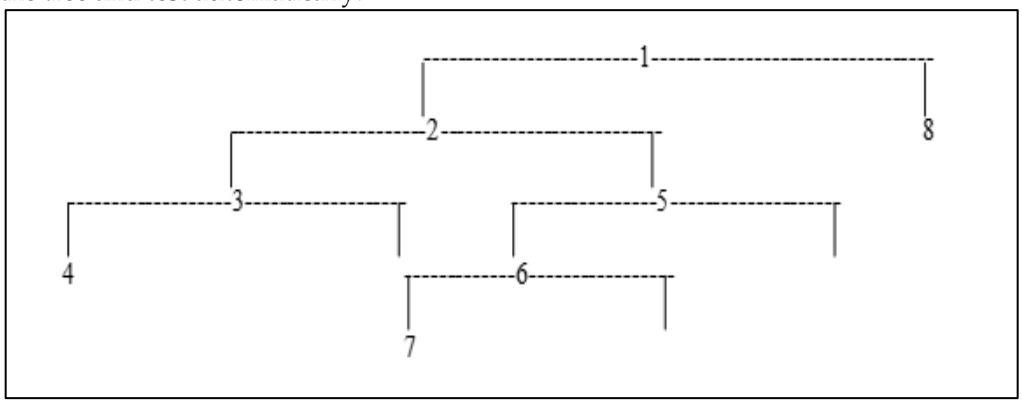

Figure 10. Classification Tree by CART

To classify new data in CART, decision trees are used because CART uses a learning sample for the building of a decision tree, and samples are divided into smaller subsets as shown in Figure 10. The author of the study Holland [48] said about decision trees are represented by a set of queries that separates the learning process. So, the tree can be constructed with the help of datasets and their dependent information, regression, or classification. Although the classification method is frequently used in CART by using historic data for the construction of a decision tree. To obtain effective results from this algorithm, first, we need to know about class distribution which has already been discussed in Table 4.

Yes, or No (inversion type) questions are only asked by the CART algorithm. Possible questions asked by CART are: Is age greater than 60? Is the room furnished? To search out the best split, the CART algorithm will find all the possible values and variables. Figure 11 shows the example of patients with multiple risk levels.

The decision tree includes hundreds of levels and numerous variables. Both numeric and categorical variables can be easily handled by CART as depicted in Figure 11.

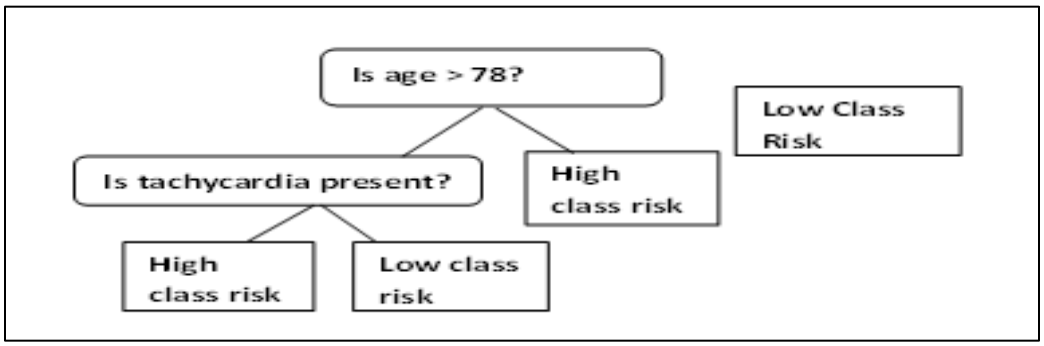

Figure 11. Patient's Classification Tree

In short, the CART methodology comprises three basic parts.

- Maximize tree construction.
- Construction of the right tree size.
- Using constructed tree, new classified data.

This is a non-parametric (inherited) approach for decision trees. Viswanathan et al. [49] said that CART is used in financial applications like Customer Relationship Management (CRM) etc. Also, they suggested that there are no assumptions about the underlying distribution. These are developed by predictor variable values. Based on a comprehensive search for all possibilities, CART recognizes separate variables. It can deal with missing variables. For non-statisticians, CART trees are very easy to understand even. Reza et al. [50] have thrown light that, Classification and Regression Trees divides only by one variable that formed unbalanced decision trees.

3.5 Java 48 (J48)

J48 is one of the decision tree algorithms. J48 is an extension of the ID3 "Iterative Dichotomiser 3" algorithm developed by the project team WEKA. Shu Hsien Liao et al. [51] research shows that C4.5 is also in strong relation with ID3 and J48. A C4.5 algorithm is an extended form of ID3. The relationship between ID3, C4.5, and J48 is mentioned in Figure 12.

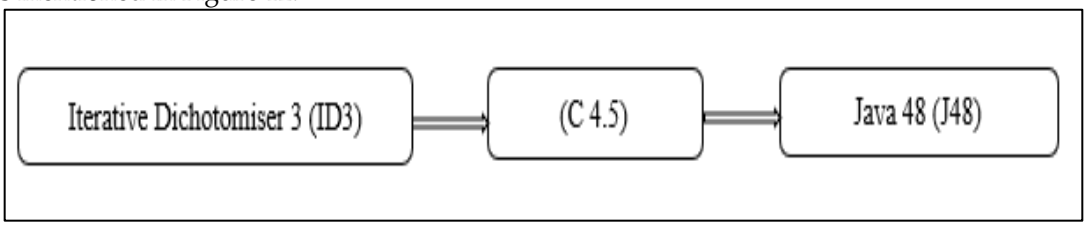

Figure 12. Origination of Java 48 (J48)

In WEKA, C4.5 is known as J48 (J stands for Java). C4.5 is also known as a statistical classifier.

In short, it can be said that J48 is an extended form of C4.5. This algorithm has multiple features like "Rules Derivation, Accounting for missing values, Value Range as well as Decision Tree Pruning". Nazmy El-messiry et al. [52] throw light and concluded that J48 is open source implementation of the C4.5 algorithm.

To decide the target value of a new sample based on multiple parameters, the decision tree is a predictive model based on machine learning. A decision tree can be expressed as the terminal or destination node which tells us the final value, different nodes of the decision tree denoted by internal nodes, and possible values identified by the branches between the nodes. J48 uses two types of variables. These are independent and dependent variables. Dependent variables are those that are predicted because their values depend upon some other attributes. Independent variables are those that provide help in predicting the value of dependent variables.

J48 decision tree algorithm follows a simple algorithm to classify new items. The first step which will take place to classify a new item is to create a decision tree based on a valued attribute from the available training data set. Moreover, it will look for other attributes with maximum or highest information gain. J48 is also

referred to as a statistical classifier because it can handle numeric and categorical data. J48 supports medium size data sets with faster training time and low expenditure cost. To create a decision tree by using the J48 algorithm, there are two basic steps which are listed below.

- The first step is to analyze the database.
- The second step is to prepare the database by accessing information.
- The information must be valid. If invalid or incorrect data is imported into WEKA, no algorithm will be able to analyze the database properly and efficiently.
- Then get started with the WEKA program.

Each leaf node in J48 demonstrates performance parameters until the data should have a possible and perfect count rate as elaborated in Figure 13. This algorithm generates rules from which it produces specific data. It gives balanced flexibility and accuracy when the goal is to gradually generalize a decision tree.

Some working steps of the J48 algorithm are discussed below.

The first case is if the instance belongs to the same class then the tree is represented by a leaf so that by labeling the same class, the leaf is returned.

By testing the attributes, for every attribute information is computed. The result from the test is also calculated which is said to be information gain.

- Then choose the best attribute selection criteria and the branch that is found on the property.
- J48 algorithm has been used to enhance the accuracy rate of data mining procedures.

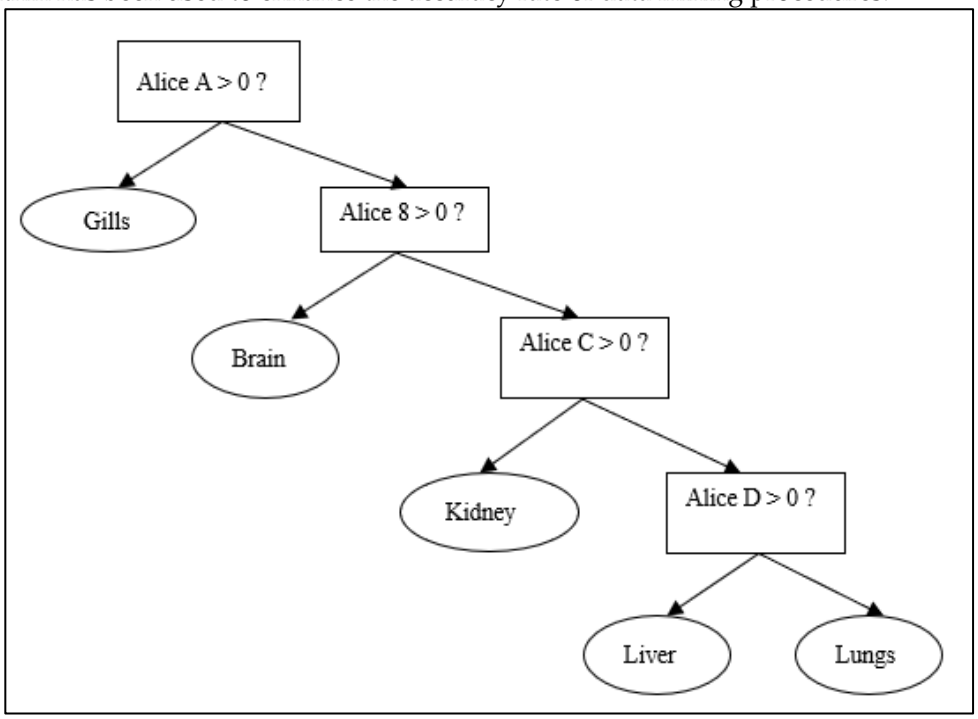

Figure 13. J48 Pruned Tree

3.6 Cross Validation Summary

This section splits the data set into various k-folds for training and testing. It also describes the training of the CART and J48 algorithms to predict the classes from given attributes. For the effective performance of the model, we train our data set by using 10-fold cross-validation which is shown in Figure 14. The data set is divided into 10 groups. To compute the average of all k-test, the model is tested and repeated. Figure 14 depicts that the data set is split into 10 groups in which 1 group is considered as test data and all other remaining are considered as training set data. Assessment outcomes are retained as 92% in Round 1, 89% in Round 2, 90% in Round 3, and 95% in Round 10.

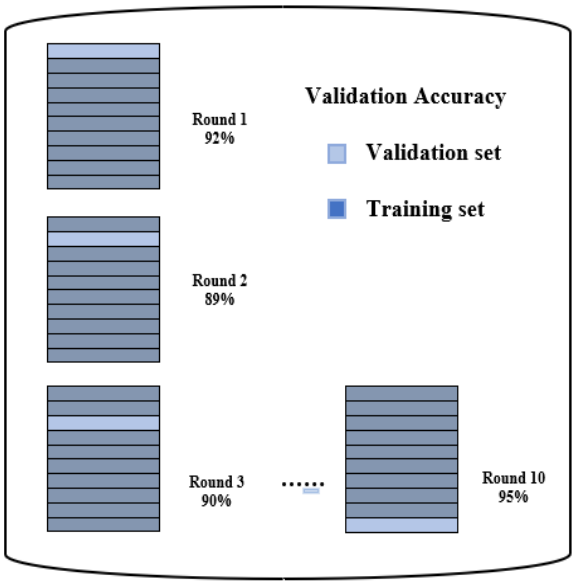

Figure 14. 10-Fold Cross Validation

3.7 Parameters for Performance Measurement

Distinctive performance measures are used to evaluate the performance of the classifier.

These Sensitivity, Specificity, Accuracy, Precision, F-Measure, ROC curve, and Confusion matrix are discussed below.

- **Sensitivity:** Test ability to authenticate those who have the disease, in other words, can be said the "True Positive" rate. Sensitivity deals with the proportion of actual positives that are correctly identified as True Positives (TP). The formula to calculate sensitivity is written below and Figure 15 shows the high sensitivity and low specificity.

  Sensitivity = TP / TP+FP **[1]**

- **Specificity:** (Mehmet Korurek et al. [53] had said that the test's ability to authenticate those who do not have the disease, in other words, can be said the "True Negative" rate. Specificity deals with the proportion of negatives that are correctly identified as negative. The formula to calculate specificity is given below.

  Specificity = TN / TN+FN **[2]**

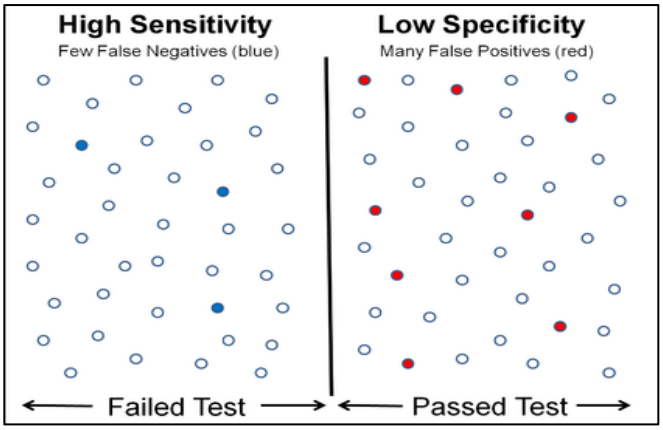

Figure 15. Sensitivity and Specificity

- **Accuracy**:

Accuracy is the basic and simple performance measure. Accuracy is the ratio of the number of valid or correct classifications to the total number of correct or incorrect classifications. The formula to calculate accuracy is written as follows:

Accuracy = TP+TN / TP+ FP+TN+FN **[3]**

- **Precision:**

Precision is used to describe what ratio/proportion of predicted true identifications were valid and correct. The formula to calculate accuracy is written as follows:

Precision = TP/TP+FP **[4]**

- **Frequency Measure (F-Measure):**

Frequency Measure is the computed weighted average of sensitivity and precision values. It is also applicable in the information retrieval domain for the approximation and estimation of classifier performance. The formula to calculate accuracy is written as follows:

F-Measure = 2*sensitivity*precision / sensitivity + precision **[5]**

- **Recall:**

The recall is used to compute the actual Positive Rate (PR). From overall positive data points, the recall rate represents the ratio of the data points that are classified as accurate and positive. The formula to calculate recall is written as follows:

Recall = TP / TP+FN **[6]**

3.8 ROC Curve

ROC stands for "Receiver Operating Characteristics". ROC is the very basic tool for diagnosing test evaluation. The ROC curve depends upon two basic parameters which are sensitivity and specificity as shown in Figure 16. ROC "Receiver Operating Characteristic" is wedged between sensitivity and specificity. The main benefit of ROC is to diagnose the accuracy and interpretation of investigation that makes available analysis. This analysis process is attached directly to cost in and normal manner. Finally, all collected parameters like sensitivity and specificity report a diagnostic and analytical accuracy that can offer tests (Mohamed Owis et al, [54].

ROC curve is also known as the "Relative Operating Characteristic" curve, as it provides a comparative analysis of two operating parameters which are True Positive rate & False Positive rate. To illustrate the performance of binary classifiers, the ROC curve provides a graphical plot as provided in Figure 16.

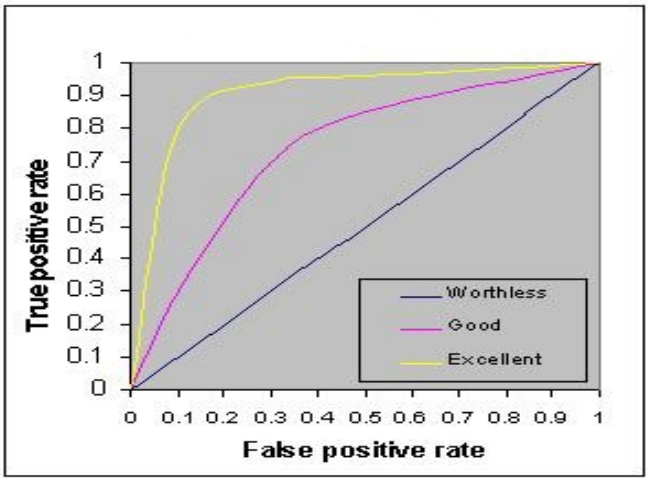

Figure 16. ROC Curve

In ROC Curve false positive rate is on X-axis and the true positive rate is on Y-axis. ROC curve can be categorized as excellent (0.90-1), good (0.80-0.90), fair (0.600.70), or fail (0.50 – 0.60) in the test.

3.9 Confusion Matrix

The confusion matrix is used as proof of the characteristics of quality management. It contains several contents that have been classified correctly for each class [54]. "True positive", "False Positive", "True Negative" and "False Negative" are computed with the help of a confusion matrix as shown in Table 6.

- **True Positives** (TP) are the number of examples properly and correctly classified to that class.
- **True Negatives** (TN) are the number of examples correctly discarded or rejected from that class.
- **False Positives** (FP) are the number of examples incorrectly classified to that class.
- **False Negatives** (FN) are the number of examples imperfectly or incorrectly rejected from that class.

Table 6. Confusion Matrix

	P′ (Predicted)	**N′ (Predicted)**
P (Actual)	Predicted True Positive	Predicted False Negative
N (Actual)	Predicted False Positive	Predicted True Negative

4. Results

In this study, J48 and CART classifiers have been applied. Table 7 provides the cross-validation summary for both classifiers. The total number of attributes is 20 which has already been discussed in Table 4. This study comprises a 10-fold cross-validation test mode that has been discussed in Figure 13 with validation accuracy from round 1 to onward round 10. Table 8 shows the class accuracy by using the J48 classifier in terms of True positive, False positive, precision, recall, frequency measure, and ROC area. The average weight has combined results for both classes with a full training-set classification model.

Table 7. Cross-Validation Summary

Instances that are correctly classified	**124 (80%)**
Instances that are incorrectly classified	3 (20%)
Kappa statistic	0.3966
Mean absolute error	0.241
Root mean squared error	0.4308
Relative absolute error	72 %
Root relative squared error	106%
Total No. of Instances	155

Table 8. Accuracy by using J48 Classifier

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
0.87	0.463	0.88	0.87	0.873	0.67	2
0.531	0.13	0.515	0.531	0.523	0.67	1
0.8	0.399	0.802	0.8	0.801	0.67	

Average Weight					

Table 8 shows the class accuracy that has been achieved by using the J48 classifier. True positive, False positive, precision, recall, frequency measure, and ROC area were obtained as 0.87, 0.46, 0.88, 0.87, 0.873, and 0.67 for class 2 respectively. In the same way for class 1, the same parameters were obtained as 0.53, 0.13, 0.515, 0.531, 0.52, and 0.67 respectively. The average weight has combined results for both classes related to the J48 classifier.

Table 9. Results using the J48 Algorithm

Instances that are correctly classified	**124 (80%)**
Instances that are incorrectly classified	31 (20%)
Kappa statistic	0.396
Mean absolute error	0.241
Root mean squared error	0.430
Relative absolute error	72%
Root relative squared error	106%
Total no. of instances	155
No. of leaves	8
Size of tree	15

Our data set contains 155 attributes as mentioned in Table 9, which gives multiple calculations that have been accessed by the J48 classifier. The number of instances that are corrected and classified by the J48 algorithm is 124 with 80%. The number of instances that are incorrect with the same algorithm is 31 and their calculated percentage is 20. Kappa statistic is an agreement of measurement between multiple values and variables that is 0.396. Other parameters like mean absolute error, root mean squared error, relative absolute error, and root relative squared error are obtained at 0.24, 0.43, 72%, and 106% respectively. The total number of leaves that have been taken by J48 is 8 with a tree size of 15.

Table 10. Confusion Matrix by J48 Algorithm

	P′ (Predicted)	**n′ (Predicted)**
P (Actual)	107	16
N (Actual)	15	17

Table 10 provides the stratified summary of the confusion matrix for the J48 algorithm that is already mentioned in Table 6 in the form of actual and predicted theory. Equations [3], [1], and [2] discussed previously, give us a deep understanding of accuracy, sensitivity, and specificity ratios obtained by the J48 classifier as 80%, 87%, and 51% which are given below. A descriptive analysis of these parameters has been already provided in the performance measurement section.

$$\text{Accuracy} = TP+TN / TP+TN + FP +FN$$
$$107+17/107+17+15+16 = 0.8 = 80\%$$

$$\text{Sensitivity} = TP / TP+FP$$
$$107/107+15 = 0.877 = 87\%$$

Specificity = TN / TN+ FN
17/17+16 = 0.515 = 51%

Table 11. Class Accuracy by CART Algorithm

TP Rate		Fp Rate	Precision	Recall	F-Measure	Roc Area	Class
0.935		0.94	0.793	0.93	0.858	0.543	2
0.063		0.065	0.2	0.063	0.095	0.543	1
Average Weight	0.755	0.757	0.671	0.755	0.701	0.543	0

Table 11 shows the class accuracy that has been achieved by using the CART classifier in terms of True positive, False positive, precision, recall, frequency measure, and ROC area. The average weight has combined results for both classes related to the CART classifier.

Our data set contains 155 attributes. Table 12 shows multiple calculations that have been accessed by the CART classifier. Correctly, classified instances by the CART algorithm are 117 (75.4%), and incorrect instances are 38 (24%). Table 12 also showed other parameters such as Kappa statistic, mean absolute error, root mean squared error, relative absolute error, and root relative squared error. No. of leaf node and tree size has been calculated as 1 for both.

Table 12. Results using the CART Algorithm

Instance	Value
Instances that are correctly classified	117 (75%)
Instances that are incorrectly classified	38 (24%)
Kappa statistic	-0.0034
Mean absolute error	0.318
Root mean squared error	0.4294
Relative absolute error	96.31%
Root relative squared error	106%
Total no. of instances	155

Table 13. Confusion Matrix by CART Algorithm

	P′ (Predicted)	n′ (Predicted)
P (Actual)	115	8
N (Actual)	30	2

Table 13 provides the summary of the confusion matrix for the CART classifier that is already mentioned in above Table 6 in the form of actual and predicted theory. Equations [3], [1], and [2] discussed previously, give us a deep understanding of accuracy, sensitivity, and specificity ratios obtained by the CART classifier as 75%, 79%, and 20% which are given below. A descriptive analysis of these parameters has been already provided in the performance measurement section.

$$\text{Accuracy} = TP+TN/\ TP+TN + FP +FN$$
$$115+2/115+2+30+8 = 0.75 = 75\%$$

$$\text{Sensitivity} = TP\ /\ TP+FP$$
$$115/115+30 = 0.79 = 79\%$$

$$\text{Specificity} = TN\ /\ TN+\ FN$$
$$2/2+8 = 0.2 = 20\%$$

5. Discussion

This section provides comparative statistics and substantial characteristics for both classifiers. Keeping in the view from Table 14 and Figure 17 given below, shows that J48 provides a Sensitivity of 87.7%, Accuracy of 80%, F-Measure of 0.80%, and Time of 0.06 sec as evaluation metrics for a skewed class data set. Conversely, the CART prevails in Sensitivity at 79%, Accuracy at 75%, F-Measure at 0.70%, and Time at 0.19 sec. Table 14 also shows the best results for the Relative Absolute Error (RAE) seized from J48 and CART as 96% and 72% respectively. Assuredly, the J48 classifier accords the best performance outcomes over CART conducive to data mining procedures.

Table 14. Performance Parameters by CART & J48 Classifiers

Algo	Accu	Sensitiv	Speci	FM	P	TT	Correct Instances	Incorrect Instances	Kappa Stats	RAE	No. of leaves	Tree Size
CART	75%	79%	20%	0.701	0.671	0.19 sec	75%	24%	-0.0034	96%	1	1
J48	80%	87.70%	51.50 %	0.801	0.802	0.06 sec	80%	20%	0.396	72%	8	15

Accu=Accuracy, Sensitiv=Sensitivity, Speci=Specificity, FM=Frequency Measure, P=Precision, TT=Time Taken, RAE=Relative Absolute Error

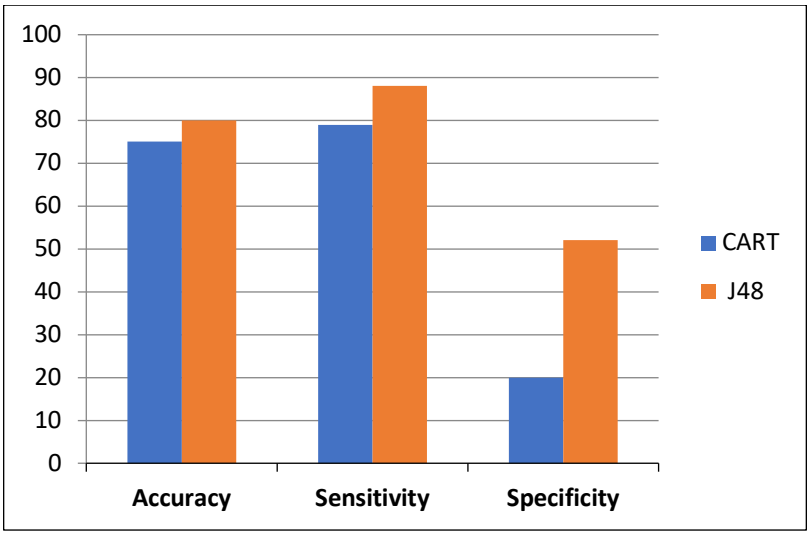

Figure 17. Comparative Analysis on the Base of Overall Accuracy

Figure 17, which is computed by performance measurements of CART & J48 algorithms, Accuracy, Specificity, and Sensitivity are clearly shown here. Accuracy by the CART algorithm is approximately 75%, sensitivity is approximately 78% and specificity is 20%. If talk about J48, its accuracy rate is approximately 80%, sensitivity is 88% and specificity is 52%. So, J48 gives the best understanding regarding performance parameters, because it can give better results for all data type attributes with the best accuracy rates and provide efficient results. It can be said that the J48 classifier has been used to enhance the accuracy rate of data mining procedures. The detailed outcomes of the performance metrics are shown in Tables [7, 8, 9, 10, 11, 12, 13, 16 (56, 57, 58, 59)] and conclusively from Tables 14 and 15, J48 yields coherent results for small to medium and skewed class data sets. This happens because of better classification, fast training time with minimum errors, better pattern explanation with minimum cost, takes less memory, minimum searching, pruning, and rules inference that lead to stable results. In comparison with CART, it gives maximum cost and takes more time to build a model as shown in Table 14. The CART classifier makes data thinner by splitting one variable which may create an unbalanced tree whereas J48 comes with multiple variables splitting criteria that formed a balanced tree as shown in Table 15. It is expected the J48 classifier will provide flexible and efficient results for large data sets as well.

Table 15. Differentiating Features for Both Classifiers

Features ⇨ Classifier ⇩	**Tree Type**	**Splitting Variable**	**Missing Values**	**Decision Tree Results**	**Type**	**Splitting Criteria**	**Type**	**Missing Values**
CART	Data thinner for each split (may create an unstable tree for large data sets)	Split only one variable	Handle missing values	May be unstable if the tree grows, increase or decrease complexity	Regression Tree Approach	Towing Criteria	Handle numerical, nominal, continuous, and Categorical values	Handle missing values
J48	Recursively splitting data criteria	Multiple variables	Handle missing values	Mostly stable	Depth First Search	Gain Ratio	Same as CART	Same as CART

Table 16. Comparative Analysis of Performance Measures of the Proposed Model

Title	**Data Set**	**Classifier**	**Accuracy**	**F-Measure**	**Recall**	**Precision**	**Time Taken (Sec)**
[56]	Iris	J48	95%	0.91	0.98	0.94	-
	Tic-Tac-Toe	J48	83%	0.79	0.86	0.99	-
	Diabetes	J48	73%	0.73	0.73	0.73	-
	Yuta- Selection	J48	66%	0.67	0.67	0.67	-
[58]	EBR Model	J48	75.40%	-	-	-	-
	ECC	J48	75.50%	-	-	-	-

	RAKEL	J48	74.50%	-	-	-	-
	EPS	J48	75.80%	-	-	-	-
[59]	Weather	CART & J48	CART (50%) J48 (50%)	0.42	0.5	0.37	0.07
	Hepatitis	CART & J48	CART (69.14%) J48 (72.57%)	0.72	0.72	0.72	0.05
	Zoo	CART & J48	CART (40%) J48 (80%)	0.92	0.89	0.92	0.06
[57]	Hepatitis	J48	J48 (79.1%)	-	-	-	-
Pro-posed Method	Hepatitis	CART & J48	CART (75%) J48 (80%)	0.80, Sensitivity (87.7)	0.95	0.8	0.06

Table 16 shows the detailed comparative analysis of performance measures of various studies and the proposed study. It can be seen that for the hepatitis dataset, J48 provided the best results.

6. Conclusions

In the present research, we aim to make the reader understand the comparative and conclusive analysis in applying machine-learning algorithms to non-linear data sets for disease prediction. The finding in our research states that J48 outperforms the CART classification algorithm and gives the best in terms of performance evaluation metrics and characteristics as discussed in Table 14 and Table 15. It is also noticed that J48 is advisable for medium and large data sets due to multifarious features as discussed in Table 15 but on the contrary, CART is convenient for small data sets. It can also be used for large data sets but with paramount time and cost. Moreover, decision trees performed well for skewed class data sets (the data analysis related to one of two classes like in the hepatitis classification problem, the patient has hepatitis is 1% so p = 1 (Yes), and the patient who does not suffer from hepatitis is 98% then p = 0 (No) [55].

Although, we used a small sample size for this examination that gives appeased results by the J48 algorithm as shown in Table 16. Moreover, if a large data set is used with resampling, threshold-moving, and clustering the abundant class techniques, it can surmise healthier results by minimizing the over-fitted and unbalanced issues too [55] as discussed in the literature review. Hence, it can be said that DT is effective for balanced classification with small to medium and skewed data sets, however, it cannot generate remarkable outcomes and limitations on imbalanced data sets. Keeping in view both classifiers' comparison hence, it can be said that J48 is strongly recommended to help in effective decision-making. This decision support system has a very great potential to be further improved in the future by using ensemble-learning algorithms that will accouterment the limitations. The former will apply a data augmentation technique to enhance the limited-size dataset before feeding it to the machine learning models. In the latter, we will experiment from the beginning with a large-scale non-synthetic dataset.

Disclosure: This work is available in Research Square as a preprint article; it offers immediate access but has not been peer-reviewed [60].

Ethics Approval: Ethics approval of this study was granted by the Human Research Ethics Committee of the National University of Sciences and Technology, Pakistan.

Availability of Data and Materials: The dataset analyzed during the study can be accessed at https://archive.ics.uci.edu/ml/datasets/hepatitis.

Authors' Contribution Statement: AJ has written the manuscript, TK reviewed and written some other main points**.** AJ and ISB developed the model and derived results using Weka. AJ edited the whole article. All authors were involved in the interpretation of results and all authors have read and approved the final version of the manuscript.

References

1. B.E. Wang, Treatment of chronic liver diseases with traditional Chinese medicine J Gastroenterol Hepatol, IEEE International Conference on Robotics and Auto-mation (ICRA) (2000), 15 Suppl, 2000, pp. 67–70.
2. N. Cheung, Machine Learning Technique/or Medica1 Analysis. B.Sc. thesis, University of Queensland, 2001.
3. Nilsson N. Introduction to Machine Learning an Early Draft of a Proposed Textbook. Stanford University: Stanford CA94305 (1998).
4. Roslina P. and Noraziah O. Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method. Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), 2010, pp.2209-2211.
5. Roslina P. and Noraziah O. Prediction of Hepatitis Prognosis Using Clinical Method and Wrapper Method. Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), 2010, pp.2209-2211.
6. Manyika, Pedro D. and Stephen M. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, (2011).
7. Sathyadevi J. Application of J48 algorithm in hepatitis disease diagnosis. Department of Computer Science and Engineering Anna University of Technology, Tiruchirapalli, Tamil Nadu, IEEE-International Conference on Recent Trends in Information Technology, (2011), ICRTIT 2011MIT, Anna University, Chennai
8. Stuart R. and Peter N. Artificial Intelligence A Modern Approach" Library of Congress Cataloging-in-Publication Data, Contributing writers: John F. Canny, Jitendra M. Malik, Douglas D. Edwards Prentice Hall, Englewood Cliffs, (2012) New Jersey 07632.
9. Chawla M. PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: A survey and comparison, Applied Soft Computing, (2011) pp 2216–2226.
10. Stephen M. Machine learning an algorithmic perspective. Massey University New Zealand, (2012) CRC Press.
11. Fahad R. Albogamy et al., Decision Support System for Survivability of Hepatitis Patients. (2022), https://doi.org/10.3389/fpubh.2022.862497
12. RT Sutton et al., An overview of clinical decision support systems: benefits, risks, and strategies for success. NPD digital medicine (2020).
13. https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96 (2019)
14. Bekir K. Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers. Journal of Science and Technology, (2016) Volume.1, Number.1 pp.49-62.
15. Mahdieh A. and Hassan Z. Automatic disease diagnosis systems using pattern recognition based genetic algorithm and neural networks. International Journal of the Physical Sciences Vol. 6(25), pp. 6076-6081, 23 October 2011 Available online at http://www.academicjournals.org/IJPS ISSN 1992 - 1950 ©2011 Academic Journals.
16. Yadevi J. Application of J48 algorithm in hepatitis disease diagnosis. Department of Computer Science and Engineering Anna University of Technology, Tiruchirapalli, Tamil Nadu, IEEE-International Conference on Recent Trends in Information Technology, ICRTIT (2016) MIT, Anna University, Chennai
17. Mahdieh A., Ba-Alwi and Houzifa H. Comparative Study for Analysis the Prognostic in Hepatitis Data Mining Approach. International Journal of Scientific & Engineering Research, (2016) Volume 4, Issue 8, pp.680-685.
18. Derya Adel N. and Hassan Z. Hepatitis Disease Diagnosis Using Hybrid Case-Based Reasoning and Particle Swarm Optimization. International Scholarly Research Network, ISRN Artificial Intelligence (2012).
19. Fadl Mutahir Mohamed A., Abdel R., Shamsollahi H., Christian J. and Clifford D. SS-SVM (3SVM): A New Classification Method for Hepatitis Disease Diagnosis". (IJACSA) International Journal of Advanced Computer Science and Applications, (2013) Vol. 4, No. 2, pp. 53-58.
20. Adel N., Hui-Ling C., and Da-You L. System based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation" Available online 18 August 2006 Digital Signal Processing 889–901 www.elsevier.com/locate/dsp.
21. M Afif, Okure O. and Samuel U. A Framework for Fuzzy Diagnosis of Hepatitis. ARTIFICIAL INTELLIGENCE | The mind behind AI (2000).
22. Sathyadevi, Breiman L. and Friedman G. Classification and Regression Trees, Wadsworth & Brooks/Cole Advanced Books & Software, (2000) Pacific Grove, CA.
23. Hui-Ling C. and Da-You L. , Fikes K. and Nilsson N. A new approach to the application of theorem proving to problem-solving. Artificial Intelligence, (1971) 2(3-4): 189-208.
24. Okure O., Holland L. Adaption in Natural and Artificial Systems: An introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan (1975) Press, Ann Arbor, MI.
25. Brieman, Viswanathan M. and Little M. Intelligent wheelchairs: Collision avoidance and navigation assistance for older adults with cognitive impairment. (2007) In Proc. Workshop on Intelligent Systems for Assisted Cognition. Rochester, NY.
26. Carson K. Leung et al., Machine Learning and OLAP on Big COVID-19 Data. In Proceedings of the 25th International Conference on Neural Information Processing Systems -Volume 2 (USA, 2016) DOI: 10.1109/Big Data 50022.2020.9378407

27. Pezhman Pasyar et al., Hybrid classification of diffuse liver diseases in ultrasound images using deep convolutional neural networks" In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Two (2021), IJCAI'11, AAAI Press, pp. 1237–1242.
28. Manjunath Varchagall1, D Sivakumar et al., Early Forecasting of Chronic Liver Disease from Liver Function Test Imbalance Datasets. Annual International Conference on Machine Learning (New York, NY, USA, Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 271.
29. Niranjan Panigrahi et al., An Expert System-Based Clinical Decision Support System for Hepatitis-B Prediction & Diagnosis. Sachi Nandan Mohanty, G. Nalinipriya, Om Prakash Jena and Achyuth Sarkar (eds.) Machine Learning for Healthcare Applications, (57–76) © 2021 Scrivener Publishing LLC
30. Elias Dritsas and Maria Trigka. Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. Article in Big Data and Cognitive Computing. September (2022) DOI: 10.3390/bdcc6030098
31. Sreenivasa Rao Veeranki and Manish Varshney. Intelligent Techniques and Comparative Performance Analysis of Liver Disease Prediction. International Journal of Mechanical Engineering 7115, Vol. 7 No. 1 (January 2022).
32. Anusha Ampavathia and T. Vijaya Saradhi. Multi disease-prediction framework using hybrid deep learning: an optimal prediction model, COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING (2021), VOL. 24, NO. 10, 1146–1168 https://doi.org/10.1080/10255842.2020.1869726.
33. Jeong Hyun Lee et al., Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network" European Radiology (2020) 30:1264–1273 https://doi.org/10.1007/s00330-019-06407-1
34. Neha Tanwar1 and Khandakar Faridar Rahman. Machine Learning in liver disease diagnosis: Current progress and future opportunities. IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012029 IOP Publishing doi:10.1088/1757-899X/1022/1/012029
35. Danny Varghese. IEEE International Conference on Robotics and Auto-mation (ICRA) (2017) https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222
36. Chirag Goyal. https://www.analyticsvidhya.com/blog/2021/05/25-questions-to-test-your-skills-on-decision-trees/
37. Milena Afeworki, https://milena-pa.medium.com/a-comparison-of-machine-learning-algorithms-knn-vs-decision-trees-d6110e08bfea (2021).
38. Vikrant A et al., Proceedings of the 9th International Conference on Foundations of Computer-Aided Process Design (2019) https://doi.org/10.1016/B978-0-12-818597-1.50019-9
39. Glen Nur Awaludin et al., Comparison of Decision Tree C4.5 Algorithm with K-Nearest Neighbor (KNN) Algorithm in Hadith Classification (2020), Journal of Emerging Trends in Computing and Information Sciences 1 https://doi.org/10.1109/ICCED51276.2020.9415796
40. Soyoung Park et al., Performance Evaluation of the GIS-Based Data-Mining Techniques Decision Tree, Random Forest, and Rotation Forest for Landslide Susceptibility Modeling (2019) Journal of Advanced Research in Computer Engineering & Technology (IJARCET) https://doi.org/10.3390/su11205659.
41. Sruthi E R. Understanding Random Forest. n 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA) (Dec 2016),
42. Slamet Wiyono et al., Comparative Study of KNN, SVM, and Decision Tree Algorithm for Student's Performance Prediction (2020) Journal of Advanced Research in Computer Engineering & Technology (IJARCET) http://dx.doi.org/10.12962/j24775401.v6i2.4360
43. A H Wibowo1, T I Oesman. The comparative analysis on the accuracy of kNN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency (2020) doi:10.1088/1742-6596/1450/1/012076
44. Mostafa Shanbehzadeh et al., Performance evaluation of selected decision tree algorithms for COVID-19 diagnosis using routine clinical data. In Proceedings of the 26th Annual International Conference on Machine Learning (New York, NY) (2021), https://doi.org/10.47176%2Fmjiri.35.29
45. https://www.analytixlabs.co.in/blog/decision-tree-algorithm/ (2022).
46. https://www.analyticsvidhya.com/blog/2021/03/how-to-select-best-split-in-decision-trees-using-chi-square.
47. Fikes K. and Nilsson N. A new approach to the application of theorem proving to problem-solving. Conference on Artificial Intelligence, (2017) https://www.analyticsvidhya.com/blog/2021/03/how-to-select-best-split-in-decision-trees 2(3-4): 189-208.
48. Holand Reza S., Mohammad A. and Shamsollahi P. A Nonlinear Bayesian Filtering Framework for ECG Denoising, IEEE Transactions on Biomedical Engineering (2017).
49. Viswanathan M. and Pei-Yuan H. Data mining techniques and applications, Expert systems with applications, Elsevier, (2015) pg.11303-11311
50. Reza Nazmy B., El-messiry and Al-bokhity B. Adaptive Neuro- Fuzzy Inference System For Classification of ECG Signals, Journal of Theoretical and Applied Information Technology (2009).

51. Shu Mehmet B., Korurek L. and Ali N. Clustering MIT–BIH arrhythmias with Ant Colony Optimization using time domain and PCA compressed wavelet coefficients, Digital Signal Processing, Elsevier (2010).
52. Nazmy Mohamed O., Ahmed H., Abou-Zied, Abou-Bakr, Youssef M. and Yasser M. Robust Feature Extraction from Ecg Signals Based On Nonlinear Dynamical Modeling, Proceedings of the 23rd Annual EMBS International Conference, Istanbul, Turkey (2010).
53. Snoey Mehmat H. and Elhaddad V. Analysis of Emergency Department interpretation of Electrocardiogram, Journal of Accident and Emergency Medicine, (1994) 11, 1994, pp 149-153.
54. M Owis and Peter N. "Artificial Intelligence A Modern Approach" Library of Congress Cataloging-in-Publication Data, Contributing writers: John F. Canny, Jitendra M. Malik, Douglas D. Edwards Prentice Hall, Englewood Cliffs, New Jersey 07632 (2012).
55. https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html
56. Singaravelan S. et al., A Study of Data Classification Algorithms J48 and SMO on different Datasets Journal of Pattern Recognition and Artificial Intelligence 31, (2016) http://dx.doi.org/10.5958/2249-7315.2016.00284.7.
57. Pezhman Pasyar et al., et a., Hybrid classification of diffuse liver diseases in ultrasound images using deep convolutional neural networks (2021) https://doi.org/10.1016/j.imu.2020.100496.
58. Mouna Hadj-Kacem et al., A multi-label classification approach for detecting test smells over java projects (2022). Digital Signal Processing, Elsevier, https://doi.org/10.1016/j.jksuci.2021.10.008.
59. Bilal Khan et al., CART, J-48graft, J48, ID3, Decision Stump and Random Forest: A comparative study. Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), (2022) pp.
60. Jameel, A., Bajwa, I. S., Ponum, M., Kousar, T., Afzaal, R., Ali, S., ... & Jameel, S. (2022). Prognosis of Hepatitis Disease Classification using Non Linear Compound Algorithms.