An Intelligent Medical Decision Support System

Ogbonnaya S. Success, LN Onyejegbu*

University of Port-Harcourt, Rivers State, Nigeria.

*Corresponding Author: Email: nneka2k@yahoo.com

Abstract

Continuous interpretations of numerous health examination results of patients are tedious and monotonous for clinicians; that are more often than not burdened by the overwhelming and seemingly intimidating number of patients they need to attend to routinely. This has culminated in a lot of fatal errors on the part of the medical practitioners which has led to the loss of innocent lives. Medical diagnosis is an important but complicated task that should be performed accurately and efficiently and its automation would be very useful. In this work, we developed an intelligent medical decision support system that would aid medical personnel in the accurate diagnosis and prediction of hepatitis A and B using C5.0 algorithms. A comparison was made between using the proposed C 5.0 Algorithm, and existing C4.5 algorithm. The simulation result shows that the Mean Absolute Error for C5.0 is 0.04 while that of C4.5 is 0.09%. The time complexity of the process for C5.0 is 0.03 seconds while that of C4.5 is 0.05 seconds. The correctly classified instance of C 5.0 is 98.05% while that of C4.5 is 94.81%. The result shows that C5.0 algorithm performs better than C4.5, in the diagnosis and prediction of hepatitis A and B. C5.0 machine learning algorithms was used to train the historical datasets of the hepatitis virus. WEKA simulation software was used to compute and simulate C5.0 algorithm. The methodology adopted for this work, is Object Oriented Analysis and Design Methodology. PHP programming language, HTML and MySQL database was used for the implementation.

Keywords: Hepatitis, C4.5 and C5.0 Algorithm, WEKA, Decision Tree, Medical decision Support system.

Introduction

Health examinations produces full and detailed information about the health condition of patients, the obtained information is needed for the management of individual and public issues of health. Examining health conditions does not only give health practitioners needed information to aid early diagnosis of diseases but also supplies the health examination receivers' proper recommendation for their self-health management. Both health practitioners and receivers of health examination derive benefit from the comprehensive information supplied, to obtain meaningful result from the examination process.

A detailed health examination package may compose of several examination items such as one on one physical examination, endoscopies, laboratory experiments and test, radiological studies, and others. The task of health science is to diagnose and prevent disease; the focus here is on diagnosis. Brause [5] opined that most health practitioners are posed with lots of task one of which is learning to diagnose in their early year of study. For them, the highest task in health evaluation is generating a comprehensive report after interpreting the result gotten from the examination. The results gathered from distributed sources of test are simultaneously and entirely interpreted to get the actual picture of health conditions of every health evaluation receiver. However, the detailed interpretation of multiple examination items is complex and error-prone for a busy clinician.

Continuous interpretations of numerous health evaluation results of patients are tedious and monotonous for health practitioners. Health practitioners have to spend much effort in screening plenty of health evaluation results to detect relatively few clinically significant findings. It is most
likely to have a reduction in the number and value in terms of quality of health evaluation result interpretations at the situation of human fatigue. Under this circumstance, health practitioners tend to simplify the procedures such as following their own experience rather than clinical guidelines to interpret clinical data and make decision. All these conditions potentially affect the most important issues that concerns health evaluations and quality control of all the health examination reports.

In a need of obtaining a practical, more accurate and effective computer-assistance for health evaluations impels the development process of the medical decision support system (MDSS) proposed in our research. This research developed an intelligent model for medical decision support system (MDSS) that would aid human medical personnel in the accurate diagnosis and proper prediction of Hepatitis cases A and B.

This was be achieved by developing an intelligent rules from hepatitis data set using SEE5.0/C5.0 decision tree algorithm, a model for intelligent medical decision support system (DSS) was developed using the developed rules, the system was implemented using PHP, HTML and MySQL as the database. The data for this research was gotten from University of California Irvine (UCI) Machine Learning Repository database.

Review of Related Work

The literature review measures the extent to which work has been done in relation to intelligent medical decision support system. Evan DA [9] describe four various levels whereby medical knowledge are structured in a clinical problem solving framework. They identified that observations are entities of evidence that are known as possibly pertinent in a difficult solving situation. No useful clinically facts were constituted.

Facets reflect general pathological descriptions such as aortic insufficiency, or categorical descriptions such as endocrine problem. They are also interim hypotheses that divide the information in the problem into manageable sub-problems and suggest possible solutions. Facets vary in their level of abstraction - from high level facets which may partition the problem space and may likely be a meaningful approximation to a candidate solution - to low level facets which may involve a more local inference that may explain one or two findings and would no advanced the problem solving process to the same extent.

Oguntimilehin A et. al. [14] disclosed that clinical diagnosis is basically categorization which permits physician to forecast using medical conditions and to decide suitable cause. This is a difficult decision procedure that encompasses large volume of ambiguity and also uncertainty organization mostly when dealing with disease of multiple symptoms. This experience builds a system that powerfully maps the connections, observations and corresponding diagnoses.

Dragulescuand Albu [8] implemented a system that is made of three important parts. First of logical inference is used to decide what type of Hepatitis virus is present for a new patient. The possibilities are B, B+D and C. After the second portion of the program was employed to understand what will the type be and the grade of hepatitis B (if the sick person is diseased with the virus, hepatitis B). This branch of the application is developed using methods from statistical inference. The third section of the application consist of the patients ill with C virus and it predicts the biological parameter; evolution during the treatment using AI Network.

Kahn R. et. al. [11] proposed neutrosophic classifier which employs Neutrosophic logic and is an addition of normally fuzzy classifier used. They compared it using the resulting parameters: like sum rules and the nature of associated functions. They proved that lengthy fuzzy classifier: Neutrosophic classifier; optimizes the believed parameters in comparison to fuzzy equivalent. Finally they established that Neutrosophic logic still in its emerging stage quiet holds the probable to be further experimented using different fields.

Ahmed [1] used neutrosophic and fuzzy C means gathering for enhancing CT liver appearance. Abdominal CT imaging was chosen and a segmentation approach was given to get their capability and accuracy.
The image is transmuted into Neutrosophic domain, this is designated with three subsets. The ratio of fact in the subset T, ratio of indeterminacy using subset 1, the ratio of inaccuracy using subset F. threshold of subset T, I, F was adapted using fuzzy C mean algorithm.

He concluded that Neutrosophy can reduce over segmentation and gives a better presentation on deafening and irregular images gotten using other approach since the technic can handle ambiguity and indeterminacy better. Ming R et. al. [12] applied neutrosophy to image processing by defining a-neutrosophic domain, which is termed by three subsets T, 1, and F. They then employ watershed algorithm to perform segmentation of the shape in the neutrosophic domain. The experiments showed that the plan method got better results when compared with the existing method.

Anil [3] in their, work titled "A soft computing paradigm for a medical data mining tool to predict risk of coronary heart events", they proposed a neuro-fuzzy soft computing data mining tool which was used for the prediction of the severity of risk in coronary heart events. The proposed system was based on medical data obtained from clinical observations performed on cardiac patients.

They used the concept of decision tree founded on 1D3 algorithm to classify the feature that contributes towards the diagnosis. The outcome is transformed into a crisp if then rules and then transformed into fuzzy rule base. A neuro fuzzy method is given to improve the fuzzy membership function which/outcomes in the improvement of the old system. The efficiency of the newly generated expert system, prediction accuracy is presented against the real life which was established to be operational.

Djain and Kimbi [7] proposed a medical diagnostic support system for the management of hypertension (MEDIAG). Their proposed system diagnoses the possibility of the ailment and its severity by applying fuzzy logic approach. MEDIAG applied forward chaining method in making inferences and the root sum squares of drawing inferences was employed to infer the data from the rules developed. The/defuzzification method engaged is centroid approach. The result obtained resembles human decision making with its capability to work from approximate reasoning and ultimately find a precise solution. It has 85% exact diagnosis.

Nicolae V et. al. [13] presented an approach that was ment to integrate computational intelligence with clinical investigation methods and knowledge for glaucoma diagnosis. The knowledge acquired is embedded in a fuzzy logic inference system. The resulting neuro fuzzy glaucoma diagnosis and prediction system is expected to lower the effort, difficulties and risk cost to this disease. Voula and [6] presented a new hybrid modeling methodology suitable for complex decision making procedures. Their work extended previous work on competitive fuzzy cognition maps for a medical decision support system by complementing them with case based reasoning methods.

The synergy of these methodologies is accomplished by the planned decision support application that is suitable to handle situations where decisions are not clearly distinct. The methodology developed was applied successfully to model and test two decision support systems, one a differential diagnosis problem from speech pathology area for diagnosis of language impairments and the other decision making choices in external beam radiation therapy.

Chrysostomos [10] presented an expert system for differential diagnosisiserythematosquamous disease incorporating decisions made by three classification algorithm: nearest neighbour classifier, naive Bayesian classifier and voting feature interval 5. Their proposed system enables doctors to differentiate six types Erythematosquamous diseases using clinical and histopathological parameters obtained from a patient-The system also gave explanation for the classification of each classifier.

Gt Venira [4] presented a comparison between Backpropagation and Naive Baye Classifiers to diagnose Hepatitis disease. The performance of proposed methods was selected for each classification tasks of
hepatitis disease. The overall accuracy of diagnosis systems were 98% and 97% respectively.

Sonawane [15], system ascertained that program based Medical Decision Support System (MDSS) will be very helpful for the medical professional due to its accurate and speedy decision making. They disclosed that forecasting the presence of heart ailment accurately, effects in lives saving. Sequential least optimization technique was use in support of vector machine (SVM). Database that exists at the Cleveland heart disease gotten from UCI repository was trained using SMO methods. The result shown made us to know that the MDSS is able to diagnosis faster the heart disease and produce a more correct result; it was also show excellent capability to predict large dataset.

According to Amin SU [2] medical errors are harmful and costly. It leads to thousands of mortality yearly. Hence, CDSS would give chances to decrease medical errors so as to enhanced patient safety. The proponents started a reasonable examination of the behaviour and performance of the six CDSS programs having diverse data mining methods for the diagnosis of heart disease. It was concluded based on results that the system cannot identify treatment or management options for patients with heart disease.

Sonawane [15] emphasized that decision methods integral in operating theatre body are a lot subjected to investigation, which occasionally indicate a far results from optimal. They additional uphold that the pending surgery lists had constantly been a general problem, partly owing to the point that the present hospital systems existing in Portuguese hospitals is deficient of a reliable decision support system factor to get better solution. A standardization and centralization of planning procedures which progresses reliability of the operating theater and handles the fragile situation of the waiting lists for surgery was developed using optimization methods and data mining.

Schizophrenia spectrum disorder was diagnosis by a decision system created by Kahn R [11].The system have four-stages: knowledge acquisition and organization, the development of computer-aided model, and evaluation of the overall system’s performance. The required knowledge is mined from an professional via interviews. The knowledge was modeled by algorithms and transmitted to a reckoning model formed using covering approach. The performance valuation is based on comparison of several diagnoses of eighty one pieces clinical between DSS and an expert. The outcomes showed a comparatively small rate of misclassification (18-34%) also an excellent performance or behaviour by the DSS with 66-82% accuracy.

Shankar F et. al. [16] proposes a prediction system for liver disease using intelligent machine learning. Researches provided various data to identify the causes for hepatitis. Here, Decision tree technique was applied to determine the structural information of tissues.C4.5 algorithm was used to generate decision tree that concentrates on some attributes such like age, sex, steroids, antivirals, etc for the diagnosis of the disease. These features helped in defining the abnormalities of the sick person with Heart ailment.

Figure 1: overall prototype Of CDSSs (Source: Lincoln 2001)
C5.0 combines several firsthand abilities such as adaptable misclassification costs. In C4.5, all faults are classified as alike; practically some classification faults are more severe than others. C5.0 allows a distinct charge to be distinct for each real class pair; if this choice is used, Then C5.0 makes classifiers to minimize anticipated misclassification rates slightly than error occurrence.

In an application that categorizes individuals as probable or not probable to "churn," C5.0 efforts to minimize the weighted projected error rate.

C5.0 is very easy to use. Several possibilities have been extended and simplified -- to support sampling and cross-validation, example -- and C4.5's programs used in generating rule sets have been fused into a distinct program.

The Windows form, See5 is very user-friendly with nice graphic interface.

C4.5 has a standard rule that induces classification rules in the form of a decision tree. As an addition of ID3, usual criteria of selecting splitting attributes in C4.5 are information gain ratio, in place of using information gain as that in ID3, information gain ratio avoids the bias of selecting attributes with numerous values.

Material and Method

In this research an intensive discussion of a more accurate intelligent medical prediction and decision support system model for diagnosis and prediction of Hepatitis by using signs and the symptoms of patients as variable predictors and also making use of SEE 5.0 decision tree algorithm for classification. The model was well implemented using real life sample data of Hepatitis disease. The data for this research was obtained from University of California Irvine (UCI) Machine Learning Repository database.

This work developed a model for an intelligent medical decision support system that would aid human medical personnel in accurate diagnosis and prediction of hepatitis B using C5.0 algorithm. It will develop a reliable guide for human physician and medical personnel decision support system. The proposed system will produce intelligent decision support system with ability to decrease cost and the time of uncertainty diagnosing hepatitis problems.

The system provides efficient decision process that involves lots of vagueness and uncertainty management mostly in a situation of multiple symptoms. Some of the benefits of the system includes that; it minimizes the time constrain for data classification (It optimizes the system’s time complexity), the proposed system uses a command of memory magnitude less than C4.5 during rule set construction, the C5.0 rule sets formation have noticeably lower error on unseen cases of various datasets, C5.0 expresses new processes for creating rule sets, and the development is substantial.

The use of appropriate method enhances effectiveness and the efficiency in all research work. We adopted the object Oriented Analysis and Design Methodology in analysis of this intelligent system for hepatitis diagnoses and treatment. This methodology features the unified modeling language (UMLs): Case diagram, Activity diagram, High Level Model and the architectural design of the newly introduces system. The used case diagram can be seen in Figure 2

![Figure 2: The use case diagram of the proposed system](image-url)
The use case diagram of the proposed system can be seen in figures 2 and consist of the following components:

- **C5.0 Algorithm**: split the sample established on the area that gives the highest information gain. Each subsample distinguished by the first split is then split again, usual; Based on a different field and the process continues reputedly till the subsample cannot be split anymore.

- **Training hepatitis dataset**: Training hepatitis dataset are the theoretical time series information of various symptoms and the treatment of hepatitis diseases.

- **Intelligent rules for hepatitis diagnoses**: Rule sets proposed in this work hold most of significant information in C5.0 decision tree however with less difficult model. The rules obtain with C5.0 algorithm in union with other nodes; will be help to properly diagnose hepatitis symptoms and treatment.

- **Likelihood of Hepatitis issues**: the intelligent rules developed are used to predict hepatitis cases and treatment.

- **Treatment of Hepatitis**: the developed intelligent rules for hepatitis diagnoses was adopt to interpret the symptoms from the client and the identify the treatment and the type of hepatitis case such as hepatitis B, C, D or acute Hepatitis.

- **Intelligent decision provision system**: the outcomes of our proposed model will provide intelligent guides for efficient medical decision for hepatitis treatment cases.

○ The activity diagram of the proposed system can be seen in Fig. 3.

The activity diagram of the planned system as display in figure 3.3 gives the individual presentation of the varying component of the proposed system. The raw hepatitis dataset is gotten from the database. Data mining model is used to process, integrate transformed and cleans the extracted raw hepatitis dataset. Then the C5.0 algorithm were adapt to thoroughly train the dataset to formulate rules using the obtained attributes and symptoms of hepatitis dataset. The entropy and information again of the features were compute and the dataset were been split using the dataset gain ratios. After the training all the features of trained dataset, an intelligent rule based were develop to diagnose hepatitis cases for efficient treatment and decision quick making. The architecture of the proposed system is shown in Figure 4.
The components of architecture of the proposed system as seen in figure 4 are as follow:

- **Hepatitis Symptoms**: the hepatitis symptoms were used as input to test and compute the proposed model.
- **C4.5 Algorithm**: is a machine learning algorithm were used to develop and efficient intelligent system to control and treat hepatitis cases.
- **Validated by Domain Expert**: the proposed intelligent system were been integrated with other component of the system by the experts.
- **Knowledge Base**: this contains the information store in the database that is been used to intelligently diagnose various cases of hepatitis cases.
- **Ruled Based System**: developed with C5.0 algorithm are used by the system to efficiently treat any type hepatitis case such as hepatitis A, B, C and acute hepatitis. This rules make prediction for individual records. Rue sets are derived from decision trees and, in a way, shows a simplified or distilled version of the information seen in the decision tree.
- **Display Advice to End User**: The outcomes of intelligent result for hepatitis treatment were displayed to the client

The high level model is shown in Figure 5
The High level model of the proposed system as depicted in figure 5 comprises of “Data gathering from Central Database”, “Data mining Model”, “Data warehouse”, and “Data mining techniques”, “Knowledge Discoveries” and “Decision Support System”.

Table 1: Training dataset for hepatitis symptoms (Source University of California Irvine (UCI) Machine Learning Repository database)

The WEKA simulation result for Information Gain Attribute Ratio is shown in Figure 6.

C5.0 model works by splitting the sample based on the field that provides the least information gain as shown in Figure 7. The information gain value field is displayed in Figure 7.
The C5.0 algorithm splits training data samples on basis of the highest information gain value field. The sample subset that is get from the former split will be split afterwards. The process will continue until the sample subject cannot be split and is usually according to other field. Finally, examine the lowest level split, those sample subsets that don’t have remarkable contribution to the model will be rejected. The knowledge denoted with decision tree could be mined and presented using IF THEN rules. The decision rules for hepatitis dataset classification is shown in Figure 8.

![Figure 7: Information gain value field](image)

The intelligence of our proposed system is well defined for proper diagnoses.

**Experiments Carried Out**

This application have been run and tested successfully.

Rulebase for hepatitis classification.
Table 2: Reveals the comparison of correctly categorized instances of C5.0 and C4.5

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correctly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5.0</td>
<td>98.05</td>
</tr>
<tr>
<td>C4.5</td>
<td>94.81</td>
</tr>
</tbody>
</table>

Table 2 shows correct classification of hepatitis datasets. It shows the comparison of correctly classified instances for C4.5, and C5.0 Algorithms. The correctly classified instances of C5.0 is 98.05% and C4.5 is 94.81%, this implies that C5.0 is more resourceful than C4.5 algorithm. The results are signified in Figure 9.

Figure 9: Graphical representation of comparison of correctly classified instances for C4.5, and C5.0.

Table 3: Comparison of time constrain of the process

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time complexity of the system (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5.0</td>
<td>0.03</td>
</tr>
<tr>
<td>C4.5</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 3 shows the comparison of time constrains of process for C4.5, and C5.0 machine learning algorithm. The time complexity of the process for C5.0 is 0.03 seconds and that of C4.5 is 0.05 seconds, this entails that C5.0 algorithm has minimum time complexity than C4.5. The results were graphically shown in Figure 11.
Table 4: Comparison of mean absolute error for C5.0 and C4.5 algorithm classification process.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Mean absolute error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5.0</td>
<td>0.04</td>
</tr>
<tr>
<td>C4.5</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 4 shows the comparison of Mean Absolute Error for C5.0 and C4.5 decision tree categorization process. The Mean Absolute Error for C5.0 is 0.04 while that of C4.5 is 0.09%. This means C5.0 algorithm has the minimum mean absolute error of classification than C4.5 algorithms. The results were represented graphically in Figure 11.

Figure 11: Comparison of mean absolute error for classification

Discussion of Results

In this work WEKA (simulator software for data mining) was used to simulate C5.0 to develop a decision support system for hepatitis diagnoses and treatment. The outcomes of the simulation were then used to create an intelligent system (decision support) for treatment of hepatitis cases. 100% of the training data was used with C5.0 algorithm to design rules and pattern to control various complicated cases of hepatitis problems.

The intelligent system was developed to interpret the symptoms of the hepatitis diagnoses and treatment. Using the simulator (WEKA) we obtain the attribute of a specific tree node using the C5.0 information pain, split information and gain ration for C5.0. The simulation result shows that the Mean Absolute Error for C5.0 is 0.04 while that of C4.5 is 0.09% (see Table 4). This implies that C5.0 algorithm has the minimum mean...
absolute error of classification than C4.5 algorithms. The time complexity of the process for C5.0 is 0.03 seconds while that of C4.5 is 0.05 seconds. This implies that C5.0 algorithm has minimum time complexity than C4.5 (see Table 3). The correctly classified instances of C5.0 is 98.05% while C4.5 is 94.81%, this implies that C5.0 is more efficient than C4.5 algorithm (see Table 2).

Conclusion
This work have successfully designed an intelligent Medical decision support system using C5.0 machine learning algorithm that will help to improve treatment of various cases of Hepatitis challenges. The developed intelligent system improved the uncertainty in the estimation of the treatment of various cases of hepatitis. It provides a reliable guide for hepatitis A and B treatment for physicians and medical personnel.

References