Intelligence based Hepatitis Diagnosis: An Empirical Study

Muhammad Shahroz Gul Qureshi, Bilal Khan, Noor Muhammad Khan
City University of Science and information Technology, Peshawar
*shahrozgul93@gmail.com, bilalsoft63@gmail.com, knoorm35@gmail.com

Abstract: *Liver disease is increasing on daily basis due to effect of drugs, viruses, alcohol and inherited diseases. Millions of people specially the young people's fall in death due to liver diseases. Hepatitis is the kind of liver disease which effect the population of all age of groups. For early diagnosis of these diseases, a blood test is required once a year. Besides clinic tests, machine learning and pattern recognition methods have been widely used for early diagnosis of hepatitis diseases in medicine by specialists. The major issue is that, which techniques is to be selected and why? Hence, this study presents the comparative analysis of different machine learning techniques for the early prediction of hepatitis. To evaluate these techniques MAE, RAE, Precision, and Accuracy measure are used. The overall result shows that QDA and NB perform well in reducing the error rate and increasing the accuracy.*

Keywords: *Hepatitis, Data Mining Techniques, Evaluation Measures*

1. Introduction

Hepatitis is a liver disease which target population of all age group (children, adults). Which has become a major challenge for pubic health care services like hospitals in diagnosis of hepatitis. If they perform accurate predictions on time for the disease it can save many people life[1] Hepatitis is the swelling in liver cells. Liver is the largest single organ whose function is to process nutrients from food and remove toxics from the body, build proteins and make bile.[2]. When someone caused by hepatitis it has very minimal symptoms which may lead jaundice, anorexia and malaise. Hepatitis has become the most important cause of chronic liver disease around the world and millions of people are targeted by this complication which is causing approximately 1.5 million deaths around the world in a year [3]. It can be cause by various reasons like viruses, chemicals, drugs, alcohol, inherited diseases, and patient own immune system that primarily attacks the liver cells[4]. For the diagnosis of this disease blood test is required once a year, clinic test, machine learning and pattern recognition method have been globally used for diagnosis of this disease [5]. For medical diagnosis machine learning and data mining techniques are used. Artificial intelligence approached to train computers think like humans, learn with experience and decision making from large amount of data and make a decision based on human knowledge and reasoning skills for better diagnosis hepatitis data mining techniques provide efficient tools for large datasets and to predict the severity of the disease [6]. Data mining techniques provides efficient tools to diagnose hepatitis from large dataset and to predict the severity of the disease[3][6]. Due to big data it is impossible for human interprets large amount of data. Therefore, computer is needed for extracting and new useful data. Classification is the main basic function that is executed by human brains where the classification phase in data mining. Human can analyze objects by using some characteristics to find out there similarities and differences[6]. However, this study focuses on the comparative analysis of existing and some of the new data mining techniques for early prediction of hepatitis disease. To benchmark the performance of these techniques some evaluation measures are selected that are MAE, RAE, Precision, and Accuracy. The rest of the paper is organized as: Sect. 2 describes the dataset and evaluation measures. In Sect. 3 data mining techniques are mentioned. Sect. 4 discusses the experimental outcomes. In Sect. 5 the overall conclusion is summarized.

2. Dataset and Evaluation Metrics

The dataset selected for this research is taken from Kaggle data repository available on (https://www.kaggle.com/harinir/hepatitis). This dataset contains 20 attributes listed in table on with 142 instances.

Table 1 List of Hepatitis Attributes

Attribute	Data Description	Attribute	Data Description
CLASS	Die, Live	SPLEEN PALPABLE	No, Yes
AGE	10, 20, 30, 40, 50, 60, 70, 80	SPIDERS	No, Yes
SEX	Male, Female	ASCITES	No, Yes
STEROID	No, Yes	VARICES	No, Yes
ANTIVIRALS	No, Yes	BILIRUBIN	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
FATIGUE	No, Yes	ALK PHOSPHATE	33, 80, 120, 160, 200, 250
MALAISE	No, Yes	SGOT	13, 100, 200, 300, 400, 500

ANOREXIA	No, Yes	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
LIVER BIG	No, Yes	PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90
LIVER FIRM	No, Yes	HISTOLOGY	No, Yes

Following evaluation metrics are used in this study:

Mean Absolute Error: Mean absolute error (MAE) is a measure of difference between two continuous variables. The MAE uses the same scale as the data stuff measured. This is known as a scale-dependent accuracy measure and therefore cannot be used to make comparisons between series using variegated scales [7]. This can be measure through the following equation:

$$MAE = \frac{\sum_{i=1}^{n} abs\,(y_i - \lambda(x_i)}{n} \quad (1)$$

Relative Absolute Error: Relative Absolute Error (RAE) is a way to measure the performance of a predictive model. It's primarily used in machine learning, data mining, and operations management. He RAE is expressed as a ratio, comparing a mean error (residual) to errors produced by a trivial or naive model. A reasonable model (The produces results of this are better than a trivial model) will result in a ratio of less than one[8]. For a single instance RAE can be calculated as:

$$RAE = \frac{\sum_{j=1}^{n} |P_{(ij)} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|} \quad (2)$$

Precision: Precision diagnosis is the well-judged and timely subtitle of each patient's health problem and remoter requires liaison of that subtitle to patients and surrogate decision-makers. Improving diagnosis in health care requires accumulation, validation and transformation of data into actionable information[9]. Precision can be calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Accuracy: Diagnostic accuracy relates to the worthiness of a test to discriminate between the target condition and health. Studies not meeting strict methodological standards usually over- or under-estimate the indicators of test performance as well as they limit the applicability of the results of the study [10]. The accuracy of a model can be found as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

3. Data Mining Techniques

The aim of data mining is to clarify the past and predict the future for analysis. Data mining helps extract information from large data sets. Classification, clustering, regression, association laws, internal detection, sequential patterns, and prediction are important data mining techniques [11]**.**

3.1. Artificial Neural Network

Artificial Neural Networks (ANN) were first created during the 1950s. ANN is a computational model dependent on organic neural system design and capacities. ANNs are viewed as non-direct measurable information demonstrating apparatuses in which ramified connections are shaped. An ANN has a few favorable circumstances yet one of the most recognize of these is the way that it can really proceeds from watching informational collections. ANN is in this way utilized as an interpretation device for wrong-headed capacities. Such kinds of systems help gauge the most financially savvy and perfect techniques for landing at arrangements while characterizing figuring capacities or dispersions. ANN utilizes information tests to land at arrangements instead of complete informational indexes, setting aside time and cash. ANNs are viewed as numerical models that are genuinely easy to improve existing advances for information examination. ANNs have three interconnecting layers. The principal layer is comprised of neurons input. These neurons transmit information to the subsequent layer, which thusly sends the yield neurons to the third layer [12].

3.2. Support Vector Machine

Support Vector Machine (SVM), originally developed by Boser, Guyon, and Vapnik (1992), Vapnik (1995), is based on the Vap- nik–Chervonenkis (VC) theory and structural risk minimization (SRM) principle (Vapnik, 1995, 1998), which is known to have high generalization performance [2]. Support vector can be utilized for design characterization. Which has multilayer perceptron's and outspread premise capacity systems. A thought that is key to the minutiae of the help vector learning numbering is the inward item portion between a help vector and the vector drawn from the information space. The help vectors are comprised of little subset of the preparation information removed by the calculation. Support vector learning calculations might be utilized to develop three sorts of learning

machines like Polynomial learning machines, Radial-premise capacity systems, Two-Layer perceptron's [3].Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is handled by ranking attributes for each class separately using a one-vs-all method and then taking from the top of each "pile" to give a final ranking[13].

3.3. J48 Decision Tree

J48 Decision Tree is an algorithm used to make a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [14]. This numbering produces the principles for the expectation of the objective variable. With the assistance of tree order numbering the vital diffusion of the information is powerfully justifiable [15]. Imagine that you have a dataset with a list of predictors or self-sustaining variables and a list of targets or dependent variables. Then, by applying a decision tree like J48 on that dataset would allow you to predict the target variable of a new dataset record [13].

3.4. Naïve Bayesian

Naive Bayesian (NB) classifier is a probabilistic statistical classifier. The term "naive" refer to a conditional independence among features or attributes. The "naive" assumption reduces computation complexity to a simple multiplication of probabilities. One main advantage of the Naive Bayesian classifier is its rapidity of use. That's because it is the simplest algorithm among classification algorithms. As a result of this straightforwardness, it can promptly deal with an informational index with numerous qualities. What's more, the gullible Bayesian classifier needs just little wattle of preparing information to create precise parameter estimations since it requires just the count of the frequencies of traits and property result matches in the preparation informational index [3]. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naïve yet the algorithm tends to perform well and learn rapidly in various supervised classification problems[14].

3.5. Composite Hypercube on Iterated Random Projection

Composite Hypercube on Iterated Random Projection (CHIRP) pulse compression process transforms a long duration frequency-coded pulse into a narrow pulse of greatly increased amplitude. It is a technique used in radar and sonar systems because it is a method whereby a narrow pulse with high peak power can be derived from a long duration pulse with low peak power. It can also reduce the hardware demands[16].

3.6. Quantitative Descriptive Analysis

Quantitative Descriptive Analysis (QDA) is one of main descriptive analysis techniques in sensory evaluation. QDA is a behavioral sensory evaluation tideway that uses descriptive panels to measure a product's sensory characteristics [17]. QDA approach has been recognized as a tool for measurement and optimization of sensory attributes of various food products[18].

3.7. Hoeffding Tree

Hoeffding Tree (HT) [19] is known as the spilling decision tree orientation. The name is derived from the Hoeffding bound that is castoff in the tree orientation. The elementary impression of HT is, Hoeffding bound delivers specific level of sureness on the finest attribute to riven the tree.

4. Results and discussion

This study focuses on hepatitis disease diagnosis. For experimental result the dataset from Kaggle data repository is selected as mentioned in Table 1. ANN, SVM, J48, NB, HT, CHIRP, and QDA techniques are applied using MAE, RAE, Precision, and Accuracy measures. The obtained results show that QDA perform well while decreasing the error rate for evaluation metrics as shown in Table 2.

Table 2 MAE and RAE Results

S No	Technique	MAE	RAE
1	QDA	0.2839	57.48%
2	NB	0.2891	58.83%
3	HT	0.2938	59.49%
4	SVM	0.3099	62.74%
5	ANN	0.3381	68.46%
6	CHIRP	0.3732	75.57%
7	J48	0.4395	88.99%

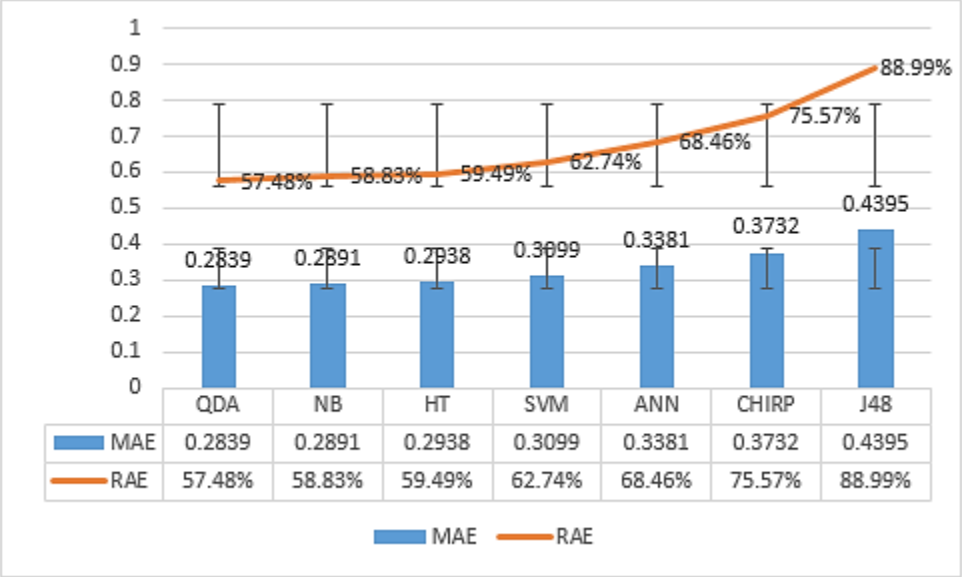

Figure 1MAE and RAE Results

While comparing the Precision and Accuracy NB perform well instead of QDA as shown in Table 3. Overall assumption shows that both QDA and NB perform well on hepatitis data.

Table 3 Precision and Accuracy Results

S No	Technique	Precision	Accuracy
1	NB	0.729	72.53%
2	QDA	0.717	71.90%
3	HT	0.719	71.12%
4	SVM	0.696	69.01%
5	ANN	0.663	66.20%
6	CHIRP	0.623	62.70%
7	J48	0.573	57.74%

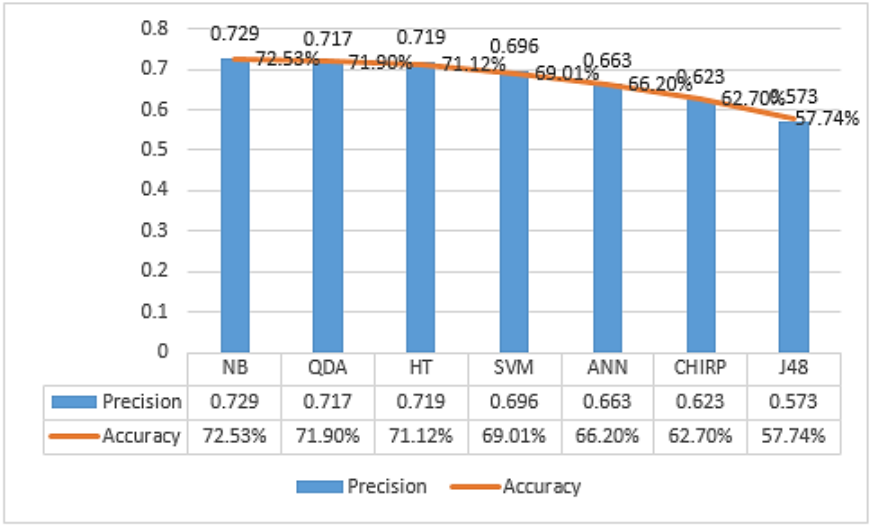

Figure 2 Precision and Accuracy Results

5. Conclusions

This study presents the comparative analysis of different classifiers. The employed classifier is tested on hepatitis dataset taken form Kaggle data repository. Experimental results show that QDA technique performance is well in reducing the error rate, while NB perform well in increasing the accuracy and precision. In future we are trying to combine the performance of QDA and NB to get high accuracy and less amount of error in the model. The hybridization of NB and QDA is the future plan of our study that can hopefully outperform good results in increasing accuracy and decreasing error rate.

6. Acknowledgements

I would like to thank you the Mr. Bilal Khan for her help and contributions, and also, I would like to thanks to USJ, for their cooperation.

References

[1] T. A. Jilani, M. Shoaib, R. Rasheed, and B. U. Rehman, "A Comparative Study of Data Mining Techniques for Hcv Patients' Data," *J. Appl. Environ. Biol. Sci*, vol. 4, no. 9S, pp. 217–223, 2014.

[2] H. L. Chen, D. Y. Liu, B. Yang, J. Liu, and G. Wang, "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 11796–11803, 2011.

[3] "Data model comparison for Hepatitis diagnosis," vol. 9359, no. 7, pp. 138–141, 2014.

[4] M. S. Bascil and F. Temurtas, "A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm," *J. Med. Syst.*, vol. 35, no. 3, pp. 433–436, 2011.

[5] Y. Kaya and M. Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease," *Appl. Soft Comput. J.*, vol. 13, no. 8, pp. 3429–3438, 2013.

[6] F. M. Ba-alwi and H. M. Hintaya, "Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach," *Int. J. Sci. Eng. Res.*, vol. 4, no. 8, pp. 680–685, 2013.

[7] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[8] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecast.*, vol. 32, no. 3, pp. 669–679, 2016.

[9] A. Belard *et al.*, "Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care," *J. Clin. Monit. Comput.*, vol. 31, no. 2, pp. 261–271, 2017.

[10] A. Amini *et al.*, "Diagnostic accuracy of tests to detect hepatitis B surface antigen: A systematic review of the literature and meta-analysis," *BMC Infect. Dis.*, vol. 17, no. Suppl 1, 2017.

[11] H. P. Ashok and G. U. Kharat, "Parallel Artificial Bee Colony Optimisation for Solving Curricula Time-Tabling Problem," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2016, no. 1, pp. 1–8, 2016.

[12] J. Dhar and A. Ranganathan, "Machine learning capabilities in medical diagnosis applications: Computational results for hepatitis disease," *Int. J. Biomed. Eng. Technol.*, vol. 17, no. 4, pp. 330–340, 2015.

[13] M. Chakraborty, R. Mehera, and R. K. Pal, "Divide-and-Conquer: An Approach to Generate All Spanning Trees of a Connected and Undirected Simple Graph," *Adv. Intell. Syst. Comput.*, vol. 3, pp. 65–84, 2014.

[14] P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 4, pp. 258–263, 2015.

[15] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, 2014.

[16] P. Song, M. W. Urban, A. Manduca, J. F. Greenleaf, and S. Chen, "Coded excitation plane wave imaging for shear wave motion detection," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 62, no. 7, pp. 1356–1372, 2015.

[17] G. R. Lloyd, S. Ahmad, M. Wasim, and R. G. Brereton, "Pattern recognition of Inductively Coupled Plasma Atomic Emission Spectroscopy of human scalp hair for discriminating between healthy and Hepatitis C patients," *Anal. Chim. Acta*, vol. 649, no. 1, pp. 33–42, 2009.

[18] R. Puri, K. Khamrui, Y. Khetra, R. Malhotra, and H. C. Devraja, "Quantitative descriptive analysis and principal component analysis for sensory characterization of Indian milk product cham-cham," *J. Food Sci. Technol.*, vol. 53, no. 2, pp. 1238–1246, 2016.

[19] N. Nahar and F. Ara, "Liver Disease Prediction by Using Different Decision Tree Techniques," *Int. J. Data Min. Knowl. Manag. Process*, vol. 8, no. 2, pp. 01–09, 2018.