International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2269-2274
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2269-2274

Hepatitis diagnosis using optimized KD-Trees and Neural Networks

Daniel Fernando Santos-Bustos[1] and Helbert Eduardo Espitia-Cuchango[2]

[1,2] *Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia*.

ORCIDs: 0000-0002-2627-9658 (Daniel), 0000-0002-0742-6069 (Helbert)

Abstract

This article uses two computational techniques to detect Hepatitis, an inflammatory disease that affects the liver. An evaluation of techniques to detect it is presented, including KD-Trees and Neural Networks. The results show that KD-Trees system had a 59% of success while the neural network system had a 72% of success rate. Also, an analysis was carried out for each result employing receiver operating characteristic curves.

Keywords: Computational intelligence, classification trees, diseases diagnosis, Hepatitis, neural networks.

I. INTRODUCTION

The help of computers as means to cooperate in the diagnosis of diseases has a wide variety of applications in health organizations. The early diagnosis of diseases helps to save lives, therefore, the development of tools that aid this end is of vital importance [1]. For this purpose, a set of classifiers was implemented to identify the presence of Hepatitis.

Hepatitis is an inflammation of the liver, most commonly caused by a viral infection. The type of Hepatitis gets its name according to the virus that produces it, namely type A, B, C, D, and E. The consumption of drugs or alcohol can also cause Hepatitis. Chronic infection with the Hepatitis B virus or the Hepatitis C virus increases the risk of liver cancer. The most common type of liver cancer is Hepatocellular Carcinoma (HCC). Other less common types of liver cancer are Intrahepatic Cholangiocarcinoma (ICC) and Hepatoblastoma. HCC has a high impact on the life of the individual, detection at early stages can decrease the number of annual deaths [2]. In this way, early diagnosis can prevent health problems that result in the prevention of viral transmission [3] and especially for diseases such as Hepatitis C for which there is no effective vaccine.

According to [4] liver fibrosis is a common middle stage of the pathological processes of chronic liver disease. Clinical intervention during the early stages of liver fibrosis delay the development of liver cirrhosis and reduce the risk of developing liver cancer.

Large populations unaware of the risk of infection often fail to meet the ideal treatment time, resulting in a poor prognosis. To cope with this, a prediction model can be used to accurately identify populations at high risk of Hepatitis infection that should take timely and appropriate medical treatment measures [5].

Given this context, it was decided to use two computational intelligence techniques that allow a classification of patients who have and do not have Hepatitis. The dataset used to implement the classifiers was taken from [6] provided by the University of California Irvine from its Open-Source repository for Machine Learning. It should be noted that this work is exploratory in order to observe the characteristics of the techniques studied to diagnose Hepatitis.

This work is divided as follows. The background that shows some applications developed to detect Hepatitis and tools used for diagnosis, then, it is shown a shallow view of the algorithms, approaches, and validation techniques used in this project, afterwards the results and analysis are presented, finally, the future work and conclusions are given.

II. BACKGROUND

Among the diseases that affect the world population Hepatitis is found to be a global population. The World Health Organization (WHO) estimates that more than 240 million people worldwide have chronic infections of Hepatitis which leads to 600 thousand deaths every year [7]. A wide variety of techniques have also been used for the detection of Hepatitis, which are variations of decision tree and neural networks, using a given number of features; the accuracy of results using a number of features is shown in [7]. A biosensor is used in [8] to detect the Hepatitis virus (HBV, HAV, HCV) immobilizing specific antibodies in the sensor element.

Supervised data extraction techniques have been successful in diagnosing Hepatitis disease through a set of data. The use of data mining techniques have been useful to develop many methods for diagnosing Hepatitis disease. Most of these methods use unique learning techniques. According to [9] the combination of the outputs of various predictors can result in improved accuracy in classification problems.

Concerning tools of diseases diagnosis, applications in the cloud such as DICOM (Digital Imaging and Communications in Medicine) have been developed in which a web application was designed for the diagnosis through a browser [10]. In [11] it is proposed a telemedicine system to diagnose diseases of the stomach. Finally in [12] they propose a new model and some key to develop CADS techniques which could be useful to share the present classifiers.

In terms of the identification of Hepatitis, a biopsy of the liver is the gold standard for the management of viral liver disease, it has drawbacks such as invasiveness and a rate relatively high sampling error. One of the most recently developed technologies, Real-time Tissue Elastography (RTE) could be a suitable imaging technology since it is non-invasive and provides accurate assessments of liver fibrosis. However, determining the stage of liver fibrosis from RTE images in a clinic is a difficult task. In this regard, [4] proposes a scheme that predicts the diagnostic stage using RTE images and multiple regression analysis, using four classic classifiers (ie, Support Vector Machine, Naïve Bayes, Random Forest and K-Nearest Neighbor) this for building a decision support system to improve the diagnostic performance of the Hepatitis B stage.

Meanwhile, [2] proposes a machine learning approach to detect HCC using a 165 patients database. In this work, a genetic algorithm is used together with the stratified cross validation method, which is applied twice. First for the optimization of parameters and then for the selection of characteristics. In the results the support vector machine (SVM) (type C-SVC) with the 2-level genetic optimizer (genetic training and trait selection) produced the highest precision.

In [9], an accurate method for the diagnosis of Hepatitis disease is proposed, taking advantage of joint learning. Iterative nonlinear partial least squares are used to perform the dimensionality reduction of the data, the self-organizing map technique for the clustering task, and the neuro-fuzzy inference system sets to predict Hepatitis disease. Decision trees are also used for the selection of the most important characteristics in the experimental data set.

In [13] a proposal is presented for the analysis of infection with the Hepatitis B virus (HBV) in human blood serum using Raman spectroscopy combined with the pattern recognition technique. Differences between normal and HBV-infected samples have been evaluated using the Support Vector Machine (SVM) algorithm. The SVM model with two different nuclei has been investigated, that is, the polynomial function and the Gaussian Radial Base Function (RBF). Furthermore, the performance of the model with each core function has also been analyzed with quadratic and least squares programming.

On the other hand, [14] constructs and compares machine learning methods in a data set of patients with the Hepatitis B virus (HBV). Models were validated on an independent HBV data set. The authors develop an easy-to-use web tool called "LiveBoost", allowing prediction models to be freely accessible for future clinical studies and applications.

Meanwhile, with the focus of establishing susceptible populations in [5], the aim is to establish models that identify high-risk populations that should be tested for Hepatitis B surface antigen. The data comes from a large examination of community-based health, which includes 97,173 individuals. A total of 33 indicators were collected, including demographic characteristics, routine blood indicators, and liver function. A synthetic limit minority oversampling technique was performed to pre-process the data and then four predictive models: extreme gradient reinforcement, random forest, decision tree and algorithms of logistic regression, were developed.

As previously mentioned, Hepatitis is linked to the development of liver cancer. On studies carried out in this regard in [15], predictive models for hepatocellular carcinoma related to chronic Hepatitis C are developed using techniques of machine learning. A data set, for 4423 patients with HCC, was investigated to identify significant parameters to predict the presence of HCC. In this study, various machine learning techniques (classification and regression tree, alternate decision tree, pruning error tree, and linear regression algorithm) were used to build HCC classification models for predicting the presence of HCC.

Another related work can be seen in [16] where the potential of radiomics is explored together with machine learning algorithms to improve the predictive precision of recurrence of hepatocellular carcinoma. A total of 470 patients were recruited from 3 independent institutions to carry out the study. In the training phase of 210 patients at institution 1, a radiology-derived signature was generated based on 3384 features extracted from the primary tumor and its periphery using an aggregated machine learning framework. Cox modeling is used to build predictive models. The models were then validated using an internal data set of 107 patients and an external data set of 153 patients from institutions 2 and 3.

III. TECHNIQUES USED

According to [17], Artificial Intelligence (AI) systems are generally evaluated in a variety of problem instances and compared to other AI systems that use different strategies. Machine Learning and supervised learning in particular are examples of this. The results of a machine learning model for an instance can be analyzed compared to other models. Although this analysis is relative to a population or distribution of models, it can give much more information than an isolated analysis. In [17] a series of experiments with a range of data sets and classification methods is performed to fully understand what parameters such as discrimination, difficulty and guesswork means for classification instances.

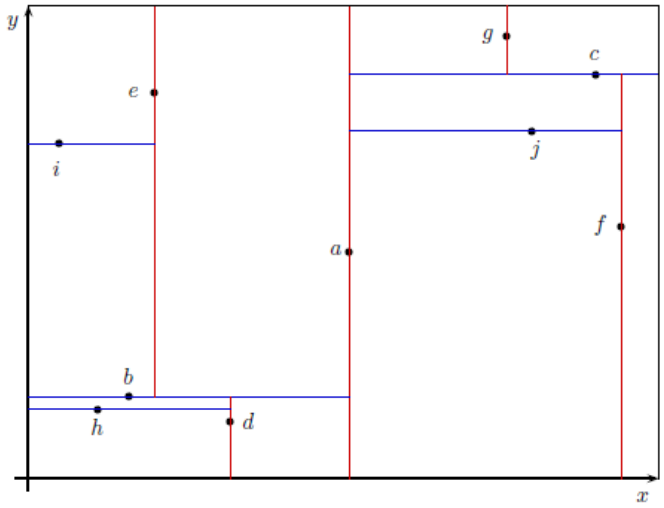
There is a variety of techniques suitable for the classification of Hepatitis given some characteristics such as Neural Networks, Decision Trees, Fuzzy Systems, among others. Neural Networks and KD-Trees were chosen to perform the classification of Hepatitis because of its ease to implement and its characteristics to fit the dataset of Hepatitis. Below is a brief review of these techniques.

III.I KD-Trees

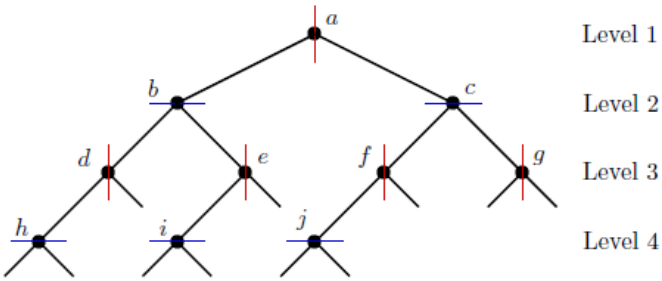
KD-Trees are a tool to partition the space of a set of dice by arranging them in a multi-dimensional binary tree, which is a specific storage structure for invested training data, in which each non-leaf node can be divided into two subspaces by a hyper-plane and recursively dividing the following subspaces

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2269-2274
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2269-2274

in two with each branch until the end of the data set. This kind of tree has many applications such as classification and finds the closest element to a given node [18]. In a KD-Tree, the divisions are made perpendicular to the plane in which the work is done, for example, Fig. 1 shows a dataset with two features (x, y) inserted in a KD-Tree as well as the lines of division of this point in a Cartesian plane. KD-Trees, also called dimensional trees, are used to sort data in a space of k dimensions, with a particular feature that are noticeably fast in consultation [19], [20]. The construction of a KD-Tree is related to obtaining the median value of the dimension under analysis, thus dividing the input data set into two groups with the same amount of data [18].



(a) KD lines that divide 10 points in a Cartesian plane.



(b) Tree representation and lines of division for a dataset with 10 points.

Fig. 1. Classification of ML algorithms.

III.II Neural Networks

The present article used a Feed Forward Network using Back Propagation as a learning algorithm. This method to determine the network parameters was applied to minimize the network error by iterative Back Propagation gradient of the error with respect to the network weights and using a numerical optimisation algorithm to reduce the network error. As the output stage of the network used was linear, it is only required to iteratively calculate the input to hidden layer weights [21].

An example of a Feed Forward Neural network is shown in Fig. 2.

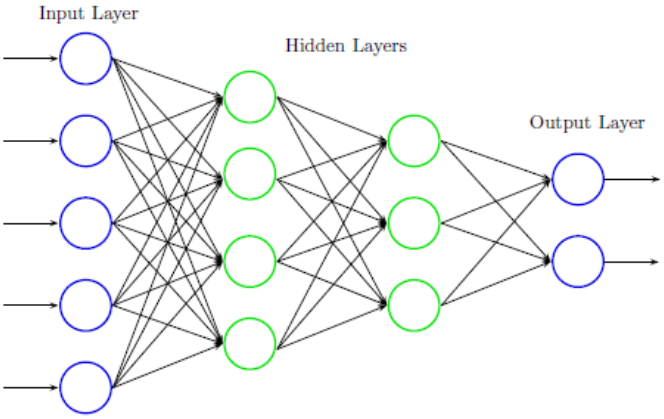


Fig. 2. Example of Feed Forward Neural Network.

IV. VALIDATION TECHNIQUES

There are three fundamental stages in which the classifiers, training, validation, and testing of the model were developed. The model is trained using the example dataset, during this stage the parameters of the classifier are adjusted, this model generates a training error which is frequently lower than the test error since the model trained is adjusted to the training data [6]. The metrics shown in this section are:

1. Confusion matrix.
2. Measure from confusion matrix.
3. ROC (Receiver Operating Characteristic) Curves.

There are many ways to evaluate classification algorithms; this section describes the measures used to evaluate the proposed classifiers. These definitions were taken from [22]. These measures must be evaluated and analyzed carefully in order to interpret the results of the model correctly.

IV.I Confusion Matrix

Some measures such as precision, recall, and sensitivity are derived from the confusion matrix, these measurements are described in [23], additionally in [24] the robustness of these measures are shown. The standard confusion matrix is shown in Table 1. The main diagonal represents the correct predictions, and the other cells the incorrect ones. In this table there are two classes P and N; the output of the predicted class is true or false.

Table 1. Example of confusion matrix.

		True / Actual Class	
		Positive (P)	Negative (N)
Predicted Class	True (T)	True Positive (TP)	False Positive (FP)
	False (F)	False Negative (FN)	True Negative (TN)
	Total	$P = TP+FN$	$N = FP+TN$

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2269-2274
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2269-2274

The measures obtained from the confusion matrix are the following:

Accuracy (*AC*): Probability that a test result be correct.

$$AC = \frac{TN+TP}{Total\ Samples} \quad (1)$$

Error Rate (*ER*): Probability that a test result be incorrect.

$$ER = \frac{FP+FN}{Total\ Samples} \quad (2)$$

Sensivity/Recall (*RE*): Probability that a test result be positive when a disease is present.

$$RE = \frac{TP}{TP+FN} \quad (3)$$

Specifity (*SP*): Probability that a test result be negative when a disease is not present.

$$SP = \frac{TN}{TN+FN} \quad (4)$$

Precision (*PR*): Probability that a test result be a positive predicted value.

$$PR = \frac{TP}{TP+FP} \quad (5)$$

F-Score (*FS*): It is a harmonic mean between Recall and Precision.

$$FS = 2 \times \frac{PR \times RE}{PR+RE} \quad (6)$$

IV.II Receiver Operating Characteristic

The receiver operating characteristics (ROC) is a curve two-dimensional graph in which the true positive rate represents the *y*-axis and false positive rate is the *x*-axis [22]. This curve has been used to evaluate many systems such as diagnostic system, medical decision-making systems and machine learning systems [25]. The characteristics of a ROC curve can be seen in Fig. 3.

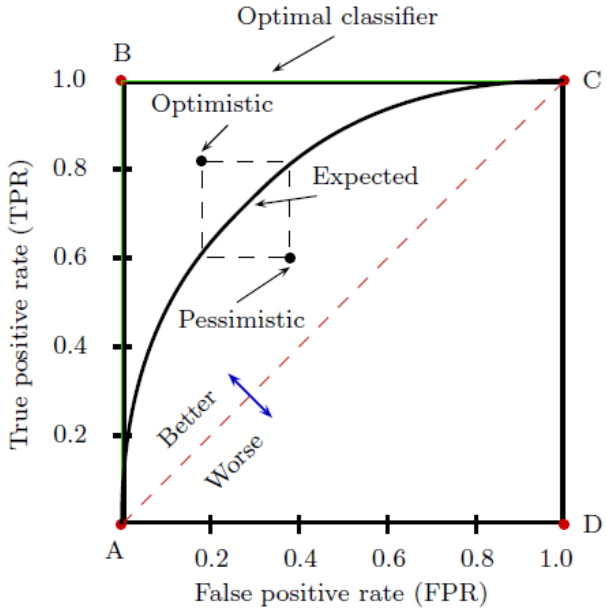


Fig. 3. Characteristics of a ROC curve [22].

V. RESULTS

This section shows the results using KD-Trees and Neural Networks considering different performance metrics.

V.I KD-Tree

For this section the KD-Tree was generated using the training dataset employing three different approaches to train and obtain the test results:

1. Cross Validation (10, 20, 30,…, 80) (Cross-*N*).
2. Leave One Out Cross Validation (LOOCV).
3. Using all data set (ALL).

Each one of the previous constructed classification trees was tested 1000 times to be able to perform statistical analyzes that would allow to approximate the real value. The results are shown in Table 2 and Table 3 as well as the ROC curves for the samples. Given the last results we compute the following measures to give comparative results.

Table 2. Standard deviation of the results using different folds in cross validation, where cross-10 means that it was used the 10% of the data to train the classifier and 90% to test (σ).

σ	*AC*	*ER*	*RE*	*SP*	*PR*	*FS*
Cross-10	0.051	0.051	0.15	0.14	0.067	0.088
Cross-20	0.043	0.043	0.12	0.1	0.055	0.067
Cross-30	0.044	0.044	0.10	0.08	0.059	0.062
Cross-40	0.046	0.046	0.09	0.07	0.06	0.06
Cross-50	0.049	0.049	0.09	0.08	0.07	0.063
Cross-60	0.054	0.054	0.1	0.08	0.08	0.071
Cross-70	0.06	0.06	0.11	0.09	0.1	0.08
Cross-80	0.07	0.07	0.13	0.11	0.12	0.10
LOOCV	0.119	0.119	0.198	0.164	0.191	0.148
ALL	0.0	0.0	0.0	0.0	0.0	0.0

Table 3. Mean value (μ) of the metrics, see Table 2 to view the standard deviation. As can be seen, using 70% of the data to train produces the best results in the classifier.

μ	*AC*	*ER*	*RE*	*SP*	*PR*	*FS*
Cross-10	0.55	0.45	0.49	0.6	0.5	0.48
Cross-20	0.56	0.44	0.5	0.6	0.51	0.5
Cross-30	0.57	0.43	0.52	0.61	0.53	0.52
Cross-40	0.57	0.43	0.52	0.60	0.52	0.52
Cross-50	0.57	0.43	0.53	0.61	0.53	0.53
Cross-60	0.57	0.43	0.52	0.62	0.53	0.52
Cross-70	**0.59**	0.41	0.54	0.63	0.55	0.53
Cross-80	**0.59**	0.41	0.54	0.64	0.55	0.64
LOOCV	**0.59**	0.4	0.54	0.65	0.56	0.53
ALL	1.0	0.0	1.0	1.0	1.0	1.0

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2269-2274
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2269-2274

The results of ROC curve can be shown in Fig. 4 observing a decreasing relation of TPR and FPR for a value of 0.63 of FPR.

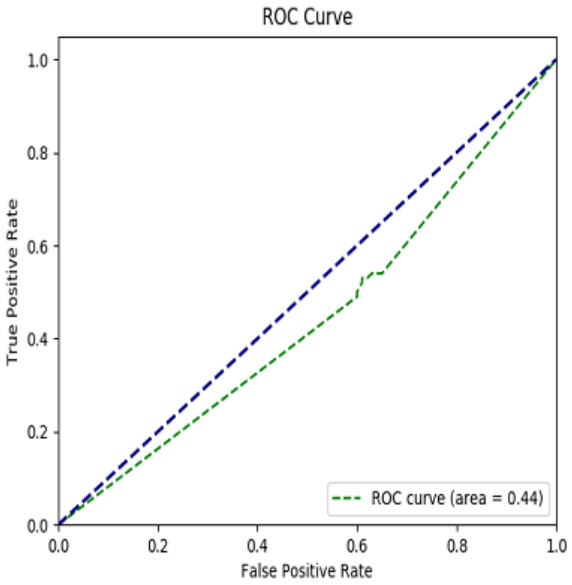


Fig. 4. ROC curve classifier using KD-Trees, showing the measurement for classification at various thresholds.

V.II Neural Networks

Since the proposed KD-Tree system produced low success rate results, it was decided to perform a neural network classifier. Supported by an experimental design, different configurations of neural networks were used. The parameters of the neural network can be seen in Table 4.

Table 4. Parameters and ranges that are used for the neural network training and test.

Parameter	Value
Hidden Layers	5
Neurons per layer	50
Loss Function	Mean Squared Error
Optimizer	Adam
Metrics	Accuracy
Activation	softmax, linear, relu, tanh

The neural network was implemented in TensorFlow, which is an interface to create and execute machine learning algorithms. It was built by Google and released under an Open Source Apache 2.0 license [26]. It allows to use run models of disease prediction in the eyes, as support to the doctor. The inputs of the neural network are the 19 features and the output represents the presence of the disease. The values of the metrics are in Table 5.

The results of ROC curve can be shown in Fig. 5 where a strong decreasing relation of TPR and FPR for a value of 0.8 of FPR is observed.

Table 5. Value metrics given the splitted data into (*k*-folds).

k-folds	*AC*	*ER*	*RE*	*SP*	*PR*	*FS*
3-fold	0.658	0.34	0.56	0.74	0.64	0.59
4-fold	0.65	0.35	0.43	0.83	0.68	0.53
5-fold	0.67	0.33	0.53	0.79	0.67	0.592
6-fold	0.67	0.33	0.6	0.73	0.65	0.62
7-fold	0.71	0.29	0.57	0.82	0.72	0.64
8-fold	0.70	0.30	0.58	0.8	0.70	0.64
9-fold	0.68	0.32	0.56	0.79	0.68	0.61
10-fold	**0.72**	0.28	0.54	0.86	0.76	0.63

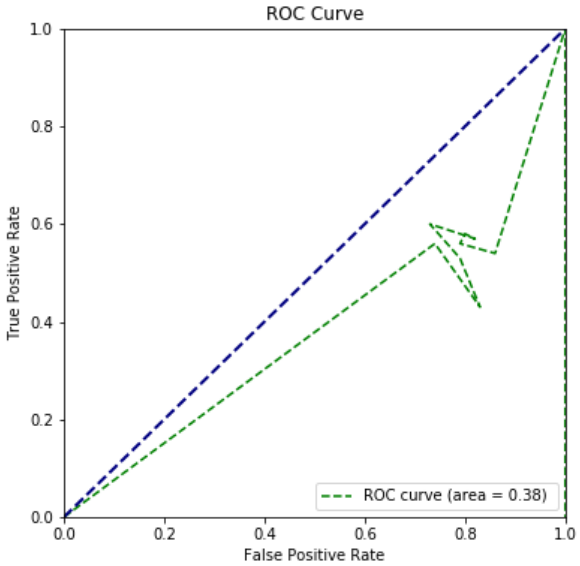


Fig. 5. ROC curve using neural network classifier. Showing the measurement for classification at various thresholds.

VI. CONCLUSIONS

The techniques of machine learning used provide an accuracy of 72% in the neural network classifier and 56% in the KD-Tree classifier. Considering the ROC the results presented showed negative results for classification of Hepatitis since the area under the curve is less than 0.5 which could mean that the usage of KD-Trees and Neural Networks have no class separation capacity whatsoever or there is no enough data in the dataset to diagnose it.

Given the state of art, it is shown that neural networks and KD-Trees are suitable classification method, then it is considered to increment the size of the dataset and implement another classifiers such as Neuro-Fuzzy and SVM that enable to increment the performance measures and to provide a suitable diagnostic system.

Finally, the current system corresponds to a model to detect Hepatitis, however, it is extensible to the implementation of other diseases as well as its implementation on devices.

REFERENCES

[1] D. Santos, H. Espitia, Detection of uveal melanoma using fuzzy and neural networks classifiers, TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(4), 2020, 2213-2223.

[2] W. Ksiazek, M. Abdar, U.R. Acharya, P. Pławiak, A novel machine learning approach for early detection of hepatocellular carcinoma patients, Cognitive Systems Research, 54, 2019, 116-127.

[3] World Health Organization (WHO), [Online accessed 2020], Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

[4] Y. Chen, et al, Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B, Computers in Biology and Medicine, 54, 2019, 116-127.

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

[6] The University of California, Irvine, Machine Learning Repository, Hepatitis Data Set (donated by Peter Turney), [Online accessed 2020], Available: https://archive.ics.uci.edu/ml/datasets/hepatitis

[7] A. Firdaus, R. Nadia, Detecting major disease in public hospital using ensemble techniques, International Symposium on Technology Management and Emerging Technologies (ISTMET), 2014.

[8] N. Yarraguntla, N. Tirumala, S. Shameem , K. Rao, Detection of hepatitis viruses (HBV, HAV, HCV) in serum using mems based bio-sensor, Second International Conference on Computing Methodologies and Communication (ICCMC), 2018.

[9] M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, E. Akbari, A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique, Journal of Infection and Public Health, 12(1), 2019, 13-20.

[10] W. Lei, W. Xi-lian, Y. Ke-hong, Design and implementation of remote medical image reading and diagnosis system based on cloud services, IEEE International Conference on Medical Imaging Physics and Engineering, 2013.

[11] X. Lu, A cooperative telemedicine environment for stomatological medical diagnosis, IEEE International Conference on Mechatronics and Automation, 2006.

[12] O. Han-bin, L. Shuai, Y. Li, H. Wen-hua, Z. Shi-zhen, Study on the new design of computer-aided diagnosis system, IEEE International Symposium on IT in Medicine & Education, 2009.

[13] S. Khan, R. Ullah, A. Khan, R. Ashraf, H. Ali, M. Bilal, M. Saleem, Analysis of hepatitis B virus infection in blood sera using raman spectroscopy and machine learning, Photodiagnosis and Photodynamic Therapy, 23, 2018, 89-93.

[14] R. Wei, J. Wang, X. Wang, G. Xie, Y. Wang, H. Zhang, C.Y. Peng, C. Rajani, S. Kwee, P. Liu, W. Jia, Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning, EBioMedicine, 35, 2018, 124-132.

[15] S. Hashem, M. ElHefnawi, S. Habashy, M. El-Adawy, G. Esmat, W, Elakel, A.O. Abdelazziz, M.M. Nabeel, A.H. Abdelmaksoud, T.M. Elbaz, H.I. Shousha, Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease, Computer Methods and Programs in Biomedicine, 196, 105551, 2020.

[16] G.W. Ji, F.P. Zhu, Q. Xu, K. Wang, M.Y. Wu, W.W. Tang, X.C. Li, X.H. Wang, Machine-learning analysis of contrast-enhanced CT radiomics predicts recurrence of hepatocellular carcinoma after resection: A multi-institutional study, EBioMedicine, 50, 2019, 156-165.

[17] F. Martíınez, R. Prudêncio, A. Martínez, J. Hernández, Item response theory in AI: Analysing machine learning classifiers at the instance level, Artificial Intelligence, 271, 2019, 18-42.

[18] V. Ayala, Cuantización Vectorial Utilizando Árboles k-Dimensionales, Proyecto de grado, Universidad de Talca, Curicó, Chile, 2018.

[19] H. Samet, Foundations of Multidimensional and Metric Data Structures, Morgan Kaufmann, 1st ed., 2006.

[20] S.S. Skiena, The Data Science Design Manual. Springer Cham, 2017.

[21] P. de Chazal, J. Tapson, A. van Schaik, A comparison of extreme learning machines and back-propagation trained feed-forward networks processing the mnist database, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[22] A. Tharwat, Classification assessment methods, Applied Computing and Informatics, (In Press), 2018.

[23] S. Shaikh, Measures derived from a 2 x 2 table for an accuracy of a diagnostic test, Journal of biometrics & biostatistics, 2(5), 2011.

[24] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management, 45(4), 2009, 427-437.

[25] K. Hajian-Tilaki, Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation, Caspian Journal of Internal Medicine, 4(2), 2013, 627-635.

[26] M. Abadi, et al, Tensorflow: Large-scale machine learning on heterogeneous distributed systems, [Online accessed 2020], Available:arXiv:1603.04467