IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 6, June 2024

Impact Factor: 5.731

Liver Diseases Prediction Using Machine Learning with Comparison Graph

Dr. Krithika. D. R[1], Dr. R. Priya[2], S. Ranjith[3]

Assistant Professor, Department of Computer Applications[1]

Professor, Department of Computer Applications[2]

PG Student, Department of Computer Applications[3]

Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India

krithika.scs@velsuniv.ac.in and priya.scs@velsuniv.ac.in

Abstract: *Chronic Liver Disease is the leading cause of global death that impacts the massive quantity of humans around the world. This disease is caused by an assortment of elements that harm the liver. For example, obesity, an undiagnosed hepatitis infection, alcohol misuse. Which is responsible for abnormal nerve function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy and there are many more. This disease diagnosis is very costly and complicated. Therefore, the goal of this work is to evaluate the performance of different Machine Learning algorithms to reduce the high cost of chronic liver disease diagnosis by prediction. In this work, we used five algorithms Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, f-1 score, and specificity. The analysis result shown the LR achieved the highest accuracy. Moreover, our present study mainly focused on the use of clinical data for liver disease prediction and explore different ways of representing such data through our analysis.*

Keywords: Liver Disease**,** obesity**,**jaundice**,** blood**,**kidney failure, Machine Learning algorithms

I. INTRODUCTION

The liver is the largest organ of the body, and it is essential for digesting food and releasing the toxic element of the body. The viruses and alcohol use lead the liver towards liver damage and lead a human to a life-threatening condition. There are many types of liver diseases whereas hepatitis, cirrhosis, liver tumors, liver cancer, and many more. Among them liver diseases and cirrhosis as the main cause of death. Therefore, liver disease is one of the major health problems in the world. Every year, around 2 million people died worldwide because of liver disease. According to the Global Burden of Disease (GBD) project, published in BMC Medicine, one million peoples are died in 2010 because of cirrhosis and million are suffering from liver cancer. Machine learning has made a significant impact on the biomedical field for liver disease prediction and diagnosis. Machine learning offers a guarantee for improving the detection and prediction of disease that has been made an interest in the biomedical field and they also increase the objectivity of the decision-making process. By using machine learning

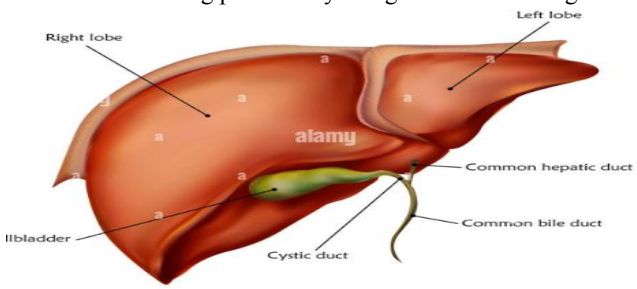

Fig1. Liver

techniques medical problems can be easily solved and the cost of diagnosis will be reduced. In this study, the main aspect is to predict the results more efficiently and reduce the cost of diagnosis in the medical sector. Therefore, we used different classification techniques for the classification of patients have liver disease or not. Six machine learning techniques have been applied including LR, KNN, DT, SVM, RF and the performance of these techniques were estimated on various perspectives such as accuracy, precision, recall, f-1 score. Moreover, the performance was compared using the receiver operative characteristic (ROC).

Python is a dynamic, high level, free open source and interpreted programming language. It supports object-oriented programming as well as procedural oriented programming. In Python, we don't need to declare the type of variable because it is a dynamically typed language. For example, x=10. Here, x can be anything such as String, int, etc. Python is an interpreted, object-oriented programming language like PERL, that has gained popularity because of its clear syntax and readability. Python is said to be relatively easy to learn and portable, meaning its statements can be interpreted in several operating systems, including UNIX-based systems, Mac OS, MS-DOS, OS/2, and various versions of Microsoft Windows 98. Python was created by Guido van Rossum, a former resident of the Netherlands, whose favorite comedy group at the time was Monty Python's Flying Circus. The source code is freely available and open for modification and reuse. Python has a significant number of users.

II. LITERATURE REVIEW

Dr. S. Vijayarani et al, delivered the e main objective of this research work is to predict liver diseases using classification algorithms. The algorithms used in this work are Naïve Bayes and support vector machine (SVM). These classifier algorithms are compared based on the performance factors i.e. classification accuracy and execution time. From the experimental results it is observed that the SVM is a better classifier for predict the liver diseases. [1Nazmun Nahar and Ferdous Ara discussed about the liver disease dataset which is select for this study is consisting of attributes like total bilirubin, direct bilirubin, age, gender, total proteins, albumin and globulin ratio. The main purpose of this work is to calculate the performance of various decision tree techniques and compare their performance. The decision tree techniques used in this study are J48, LMT, Random Forest, Random tree, REPTree, Decision Stump, and Hoeffding Tree. The analysis proves that Decision Stump provides the highest accuracy than other techniques. [2]. A.K.M SazzadurRahman, et al, found the accuracy 75%, 74%, 69%, 64%, 62% and 53% for LR, RF, DT, SVM, KNN and NB. The analysis result shown the LR achieved the highest accuracy. Moreover, our present studies mainly focused on the use of clinical data for liver disease prediction and explore different ways of representing such data through our analysis. [3]. M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi discusses liver diseases are becoming one of the most fatal diseases in several countries. In this thesis, liver patient datasets are investigated for building classification models in order to predict liver disease. This thesis implemented a feature model construction and comparative analysis for improving prediction accuracy of Indian liver patients in three phases. In first phase, min max normalization algorithm is applied on the original liver patient datasets collected from UCI repository. In liver dataset prediction second phase, by the use of PSO feature selection, subset (data) of liver patient dataset from whole normalized liver patient datasets is obtained which comprises only significant attributes. 16 Third phase, classification algorithms are applied on the data set. In the fourth phase, the accuracy will be calculated using root mean square value, root mean error value. J48 algorithm is considered as the better performance algorithm after applying PSO feature selection. Finally, the evaluation is done based on accuracy values. Thus outputs shows from proposed classification implementations indicate that J48 algorithm performances all other classification algorithm with the help of feature selection with an accuracy of 95.04%. [4]. MafazalyaqeenHassoon; Mikhak Samadi Kouhi discusses a new concept which is called Medical Data Mining (MDM). Indeed, data mining methods use different ways such as classification and clustering to classify diseases and their symptoms which are helpful for diagnosing. This paper introduces a new method of liver disease diagnosis to help doctors and their patients in finding the disease symptoms and reduce a long time of diagnosing and prevent deaths. The proposed method will optimize the rules released from Boosted C5.0 classification method with the Genetic Algorithm (GA), to increase the diagnosis time and accuracy. So instead of using an evolutionary algorithm for producing rules, the

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

genetic algorithm is used for improving and reducing rules of another algorithm. We show that our proposed approach has better performance and throughput in comparison with other work in the field [5].

III. METHODOLOGY

DATASET COLLECTION

Dataset collected from the UCI Machine Learning Repository. In addition, the original dataset. This dataset consists of liver patient's data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 11 parameters whereas we choose 10 parameters for our further analysis and 1 parameter as a target class.

PREPROCESSING

We analyzed liver patient's data whereas samples are liver patient and samples are non-liver patients. Moreover, from the liver patient's dataset are male samples and female samples were taken for analysis. The heatmap appear to have some correlated parameters. Some of these columns have a low correlation. Therefore, we omitted some of the features for better prediction of liver disease.

DATA VISUALISATION

Data visualization is the representation of data through use of common graphics, such as charts, plots, info graphics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

IMPLEMENTATION

Logistic regression can deal with any number of numerical as well as absolute factors. Strategic Regression processes the connection between the element factors by surveying probabilities (p) utilizing an underlying logistic function. Random forests or random decision forests are an ensemble learning technique for classification, regression and different assignments that works by developing a huge number of decision trees at training time and yielding the class that is the method of the classes Decision Tree calculation has a place with the supervised learning algorithms. In contrast to other supervised learning algorithms, a decision tree algorithm can be utilized for taking care of regression and classification issues as well. SVM is a supervised learning calculation. It can utilize for both grouping or relapse issues however generally it is utilized in characterization issues. KNN is one of the most fundamental occasion-based classification algorithms in Machine Learning. A KNN sorts an example to the class that is most decided among K neighboring. K is a limitation for adjusting the classification algorithms. PREDICTION We considered different analyses to examine the five-machine learning classifier for the classification of liver disease dataset. In terms of accuracy, LR achieved the highest accuracy of 75%. With respect to precision, LR achieved the highest score 91% and . When considering the sensitivity, SVM achieved the highest value 88% and KNN obtained the worst 76%. Logistics Regression was also the best performer in terms of f1 measure 83% and when considering specificity DT achieved the highest value 48% and LR the lowest 47%. According to compare these measurement criteria LR classification technique is more effective than the other classifiers for predicting chronic liver disease.

IV. DATA AND IMPLIMENTATION

Age	Gender	Total_Bilirubin	Direct_Bilirubin	aline_Phosphot	Aminotra	Aminotr	otal_Protie	Albumin	min_and_Globulin_I	Target
65	Female	0.7	0.1	500	20	34	5.9	1.6	0.37	2
62	Male	10.9	5.5	98	35	31	6	3.2	1.1	1
62	Male	7.3	4.1	245	48	49	6.4	3.2	1	1
58	Male	1	0.4	284	29	32	6.8	3.4	1	1
72	Male	3.9	2	216	21	24	7.3	4.4	1.5	2

Fig.2 Data set

IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 6, June 2024

Impact Factor: 5.731

Fig2 shows the sample data set and Fig 3 explores data distribution of numerical data. Fig 4 shows the count of male and female in data set. Fig 5 ,6, 7,8,9 ROC curve for respective machine learning model.

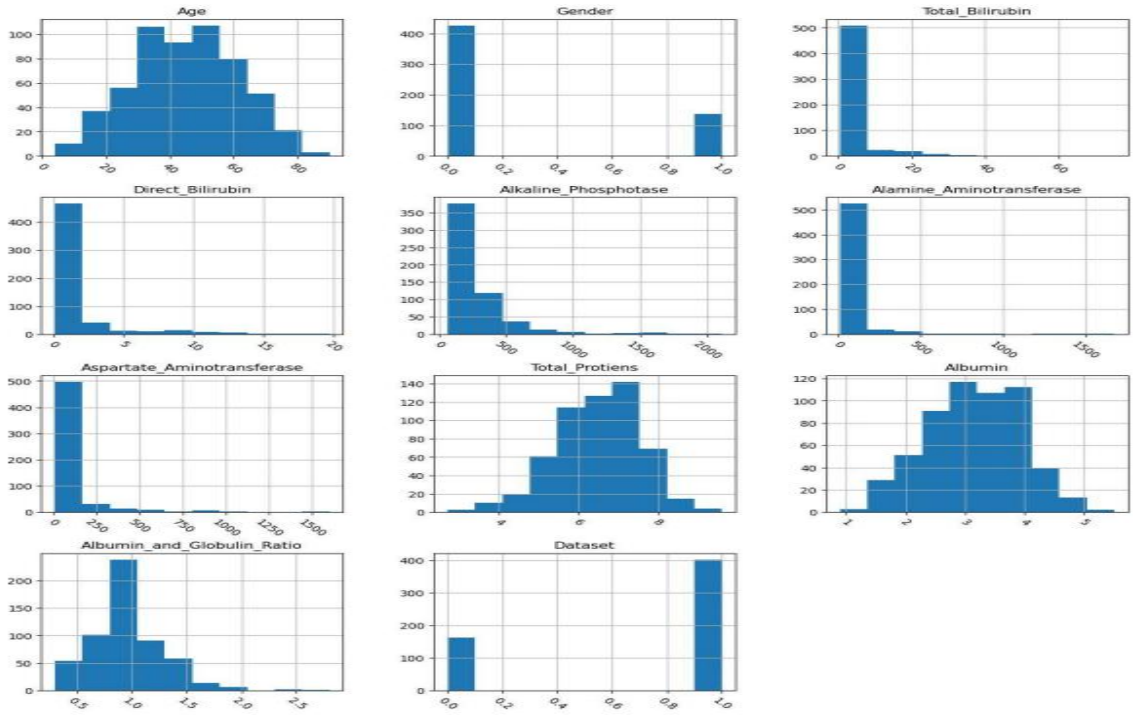

Fig 3 Distribution of numerical Features

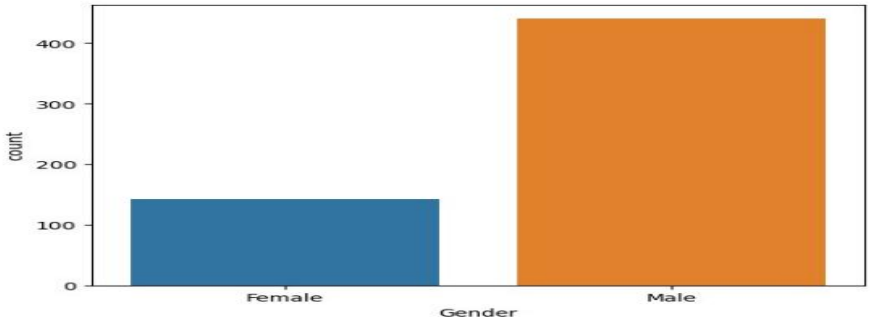

Fig 4 Count of male and female patient

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

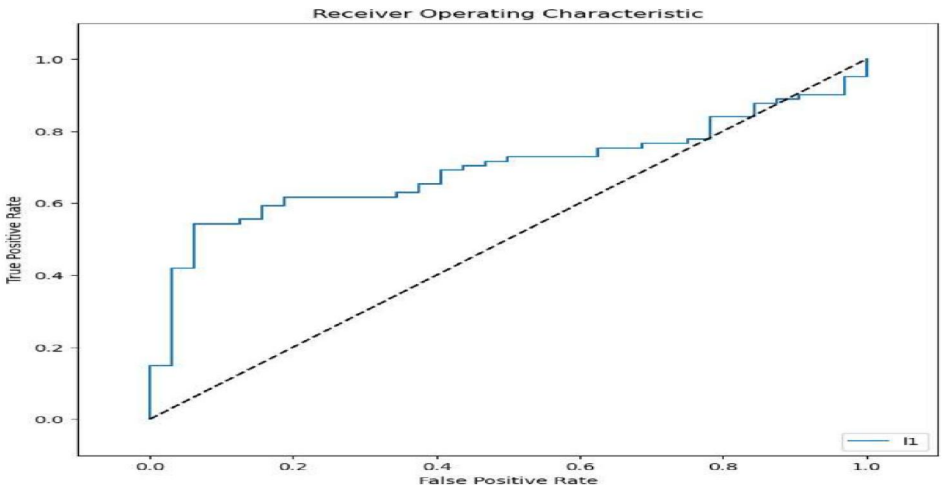

Fig. 5 ROC in Logistic Regression

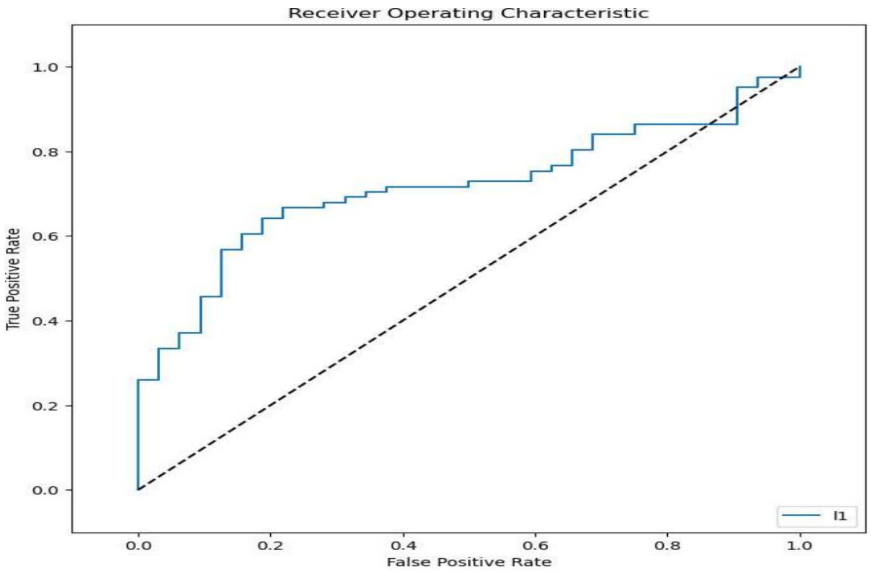

Fig 6. Random forest ROC

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

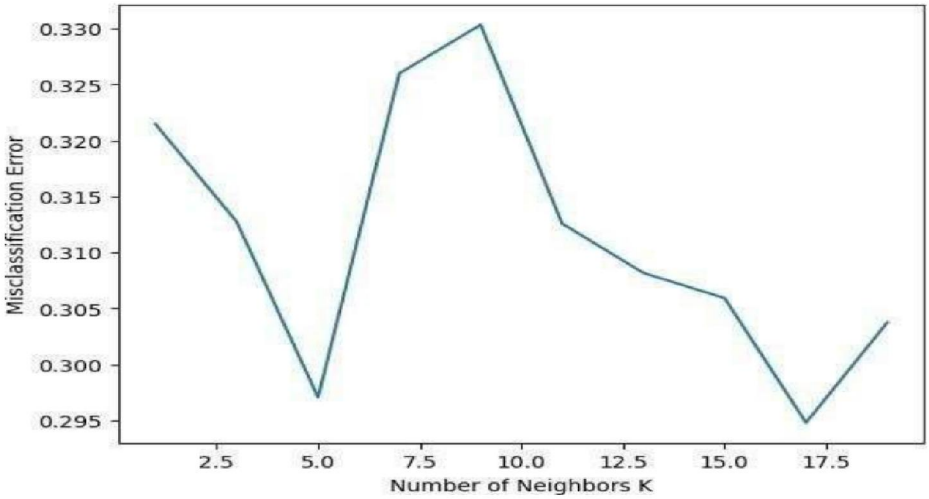

Fig 7 KNN ROC

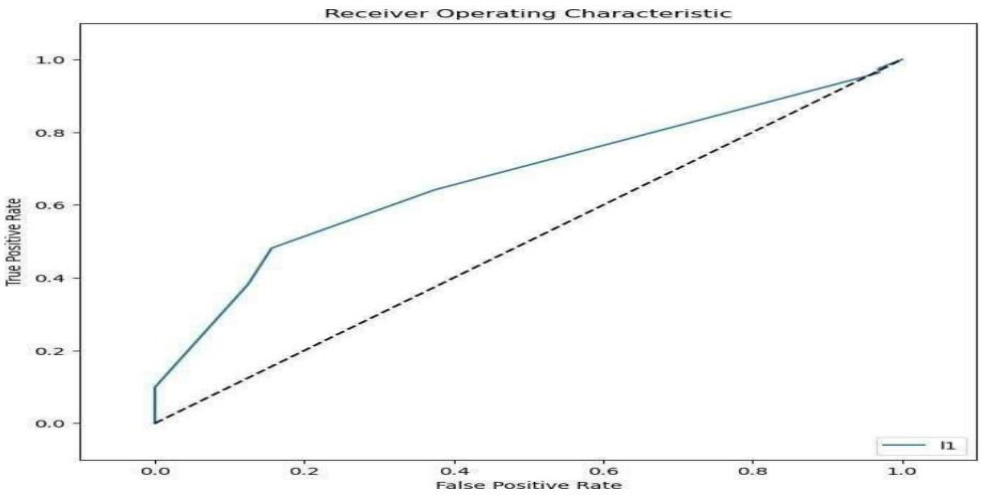

Fig 8 SVC ROC

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

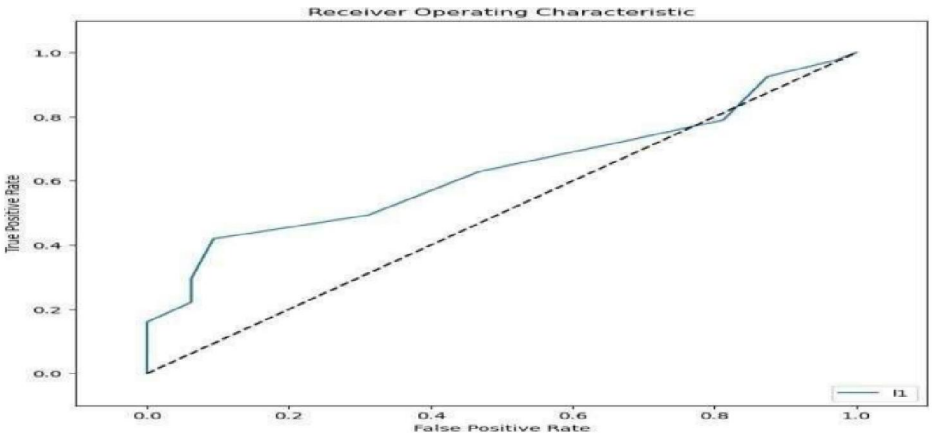

Fig 9. Decision Tree ROC

V. DISCUSSION

Logistic regression can deal with any number of numerical as well as absolute factors. Strategic Regression processes the connection between the element factors by surveying probabilities (p) utilizing an underlying logistic function. Random forests or random decision forests are an ensemble learning technique for classification, regression and different assignments that works by developing a huge number of decision trees at training time and yielding the class that is the method of the classes Decision Tree calculation has a place with the supervised learning algorithms. In contrast to other supervised learning algorithms, a decision tree algorithm can be utilized for taking care of regression and classification issues as well. SVM is a supervised learning calculation. It can utilize for both grouping or relapse issues however generally it is utilized in characterization issues. KNN is one of the most fundamental occasion-based classification algorithms in Machine Learning. A KNN sorts an example to the class that is most decided among K neighboring. K is a limitation for adjusting the classification algorithms. PREDICTION We considered different analyses to examine the five-machine learning classifier for the classification of liver disease dataset. In terms of accuracy, LR achieved the highest accuracy of 75%. With respect to precision, LR achieved the highest score 91% and . When considering the sensitivity, SVM achieved the highest value 88% and KNN obtained the worst 76%. Logistics Regression was also the best performer in terms of f1 measure 83% and When considering specificity DT achieved the highest value 48% and LR the lowest 47%. According to compare these measurement criteria LR classification technique is more effective than the other classifiers for predicting chronic liver disease.

VI. CONCLUSION

The principal part of this work is to make an effective diagnosis system for chorionic liver infection patients utilizing five distinctive supervised machine learning classifiers. We researched all classifiers execution on patient's information parameters and the LR classifier gives the most elevated order exactness 75% dependent on F1 measure to predict the liver disease. From now on, the outperform classification procedure will give for the decision support system and diagnosis of chronic disease. The application will have the option to predict liver infection prior and advise the wellbeing condition. This application can be surprisingly gainful in low-salary nations where our absence of medicinal foundations and just as particular specialists. In our study, there are a few bearings for future work in this field. We just explored some popular supervised machine learning algorithms; more algorithms can be picked to assemble an increasingly precise model of liver disease prediction and performance can be progressively improved. Additionally, this work likewise ready to assume a significant role in health care research and just as restorative focuses to anticipate liver infection.

REFERENCES

[1] Dietterich TG.(2000), Ensemble methods in machine learning. In: Proceedings of Multiple Classifier System, vol. 1857, Springer (2000), pp. 1–15.

[2] Zhi-Hua Zhou, Ensemble Methods: Foundations and Algorithms, Machine Learning & Pattern Recognition Series, 2012.

[3] Yongjun Piao, Minghao Piao, Keun Ho Ryu, Multiclass cancer classification using a feature subsetbased ensemble from microRNA expression profiles, Computers in Biology and Medicine 80 (2017) 39– 44.

[4] ManjeevanSeera, Chee Peng Lim, A hybrid intelligent system for medical data classification, Expert Systems with Applications 41 (2014) 2239– 2249.

[5] Michael J Sorich. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009;47:53—62.