NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

Predicting liver Disease Using Support Vector Machine and Logistic Regression classification Algorithm

***Osaseri R.O*[1] *& Usiobaifo A.R.*[1]**

Department of Computer Science, Faculty of Physical Sciences, University of Benin, P.M.B. 1154, Benin City, Nigeria
Email: roseline.osaseri@uniben.edu, rosemary.usiobaifo@uniben.edu

Article Info

Keywords: *liver disease, SVM, Logistic Regression.*

Received 12 August 2024
Revised 3 November 2024
Accepted 4 November 2024
Available online 7 December 2024

https://doi.org/10.5281/10.5281/zenodo.14302631

ISSN-2682-5821/© 2024 NIPES Pub.
All rights reserved.

Abstract

The liver plays a crucial role in various bodily functions, including protein production, blood clotting, and the metabolism of cholesterol, glucose, and iron. Early prediction of liver disease is vital for saving lives. In this study, machine learning algorithms were employed to predict liver disease, specifically Support Vector Machine (SVM) and Logistic Regression (LR), the primary aim is to provide an accurate, efficient, and non-invasive tool for early diagnosis and risk assessment of liver diseases. This allows for timely intervention, improved patient outcomes, and helps healthcare professionals make informed decisions. The dataset used was obtained from UCI (579 records). The models were trained with 405 samples (70%) and tested with 174 samples (30%). Key results showed that Logistic Regression outperformed SVM, achieving the highest accuracy of 97.24% and precision of 98%. SVM achieved an accuracy of 95.55% and a precision of 97%. Both models exhibited strong recall at 98% which indicates that both models are good for prediction of liver disease. The models' convergence rates were 90 epochs for LR and 4750 epochs for SVM, indicating that LR converges much faster. These findings imply that Logistic Regression not only provides better predictive performance but also converges more efficiently, making it a more suitable algorithm for liver disease prediction.

1.0. Introduction

The liver, an exocrine gland located below the diaphragm on the right side of the abdomen, is responsible for various vital functions, including bile secretion for digestion, blood purification, regulation of blood toxins, clearing bilirubin, metabolism, and converting harmful ammonia to urea. The liver plays an important role in many bodily functions from protein production and blood clotting to cholesterol, glucose (sugar), and iron metabolism. It has a range of functions, including removing toxins from the body, and is crucial to survival. The loss of those functions can cause significant damage to the body. When a liver is infected with a virus, injured by chemicals, or under attack from own immune system, the basic dangers are the same. That liver will become so damaged that it can no longer work to keep a person alive.

Diseases of the liver which last more than six months are considered chronic liver diseases. It involves the destruction of the liver parenchyma cells leading to fibrosis and cirrhosis. Chronic liver disease is an asymptomatic progressive disease and mostly fatal. They can be categorized as: (a) Viral diseases which include hepatitis B and C, Cytomegalovirus, Epstein Barr. (b) Alcoholic liver disease and drug induced liver disease from Methotrexate, Amiodarone, Nitrofurantoin and others (c) Metabolic diseases which include non-alcoholic fatty liver disease, Hemochromatosis, Wilson's disease (d) Autoimmune disorders which include autoimmune hepatitis Primary biliary cholangitis (primary biliary cirrhosis), primary sclerosing cholangitis [1]. Liver disease caused by hepatotrophic viruses imposes a substantial burden on health care resources. Persistent infections from hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus result in chronic liver disease. The most basic classification of liver disease is acute and chronic. The definition of acute liver disease is based on duration, when the history of the disease does not exceed six months. Viral hepatitis and drug reactions account for the majority of cases of acute liver disease. Liver disease is also referred to as hepatic disease. Usually nausea, vomiting, right upper quadrant abdominal pain, fatigue and weakness are classic symptoms of liver disease. Symptoms of liver patient include jaundice, abdominal pain, fatigue, nausea, and vomiting, back pain, abdominal swelling, and weight loss, fluid in abnormal cavity, general itching, pale stool, enlarged spleen and gallbladder [2]. Symptoms of liver disease can vary, but they often include swelling of the abdomen and legs, bruising easily, changes in the color of your stool and urine, and jaundice, or yellowing of the skin and eyes. Sometimes there are no symptoms. Tests such as imaging tests and liver function tests can check for liver damage and help to diagnose liver diseases.

Liver disorders have increased rapidly and it is considered to be a fatal disease in many countries like Egypt and Nigeria. The progressive increase in the cost for healthcare in recent decades is expected to continue, in fact accelerate. According to the Office for National Statistics in the United Kingdom, liver disease is now the fifth most common cause of death after heart disease, stroke, chest disease and cancer [3]. Early prediction of liver disease is very important to save human life by taking proper steps to control the disease. Thus, this study decided to employ the use of machine learning algorithm for the prediction of liver disease.

The applications of Machine Learning techniques nowadays have become very much important in the healthcare sector for the prediction of disease from the medical database. Many researchers and companies are leveraging machine learning to improve medical diagnostics. Among different machine learning techniques, classification algorithms are widely used in predicting diseases. For this research, the main aim is to predict liver disease using two classification algorithms. The algorithms used for this study are Logistic Regression and SVM.

1.1. Related work

[4] Carried out a study on classification of liver and non-liver disease dataset. Pre-processing method was used to clean the data for effective classification, after cleansing the data, 15 attributes of real medical data were collected from the dataset. C4.5 and Naive Bayes were the two algorithms used in the study. [5] did a study “Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome” the research focuses on the performance analysis of ten various and well-known ML classification algorithms on two different liver disease datasets taken from the UCI ML repository and GitHub repository. The classification algorithms include average one dependency estimator (A1DE), multilayer perceptron (MLP), NB, K-nearest neigh(KNN), SVM, composite hypercube on iterated random projection (CHIRP), decision tree (CDT), forest by penalizing attributes (ForestPA), decision tree (J48), and random forest (RF). Result from the research shows that; on UCI dataset, SVM produces better results for recall and G-measure assessment measures. On the contrary, on the dataset taken from the GitHub repository, SVM performs better in terms of increasing accuracy as well as precision and F-measure. The researchers concluded that SVM is the progressive tool with thoroughgoing classification algorithms surrounded in statistical learning theory. [6] did a work on, different classification algorithms namely Logistic Regression, Support Vector Machine and K-Nearest Neighbor; for liver disease prediction. The comparison of all these algorithms was done based on classification accuracy which is found through confusion matrix. From the experiment, Logistic Regression and K-Nearest

Neighbor have the highest accuracy but logistic regression have the highest sensitivity. [7], in their study they analyses the data related to Liver Disorder with the help of Naive Bayes, Decision Table, and J48. However, attributes like case history of the patient, diabetes, smoking, obesity, alcohol intake, and smoking were used. Based upon the given database it was concluded that male people are having more liver disorder than the females. Age group of 35-65 is mostly affected and out of these 26% people are having the disorder because of alcohol, smoking contributed to 22% of people, obesity, and diabetic of 4 & 5 percent respectively.

[8] in their proposed work the researcher have done classification of the liver patient data using the algorithms like Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron and Random Forest. The data from the UCI repository which is afforded by Center of Machine Learning and Intelligent Systems was used. After completion of their three-phase analysis, the Random Forest Algorithm came out best with accuracy of 71.87percentage. [9] proposed the concept of diagnoses of liver disease. Various classification algorithms were used such as Naïve Bayes, Ada Boost, J48, Bagging and Random Forest. These algorithms were further compared based on the parameters such as Accuracy, Error rate and so on. Also, Preprocessing technique was utilized to divide the data into two groups- liver patients and non-liver patient that was accomplished using K means clustering algorithm. Further, the clustered dataset was applied to the various classification algorithms. The implementation of the different classification algorithms was performed using the Weka Tool. The overall comparison was done between Naïve Bayes, Ada Boost, J48, Bagging and Random Forest algorithms. After the comparison was performed, the comparative study showed that the Random Forest gave the better results as compared to the other algorithms. [10] presented a novel approach to detect the liver disorders of patients at an early stage. Separation of points by planes algorithm was used to distinguish healthy patients from the unhealthy patients and assisted in the diagnoses of liver disorder. The Separation algorithm was deployed to classify the functional data of liver. The data was collected from a hospital in Hyderabad. Thus, separation algorithm could diagnose the liver disorder with the accuracy of 85.1% and the total time taken for completion of training is 1 second and testing is 1 second. [11] presented an approach of diagnoses of liver disorder through an analysis of liver disorder datasets. The main focus of the research was to help the physicians with the medical decision making process. Several algorithms were compared on various parameters such as Naive Bayes, ANN, ZeroR, IBK, VFI, J48 and Multilayer Perceptron. The algorithms were implemented using the Weka tool and dataset was collected from UCI Repository. The experimental results showed that Multilayer Perceptron gave the better classification results as compared to other algorithms. Thus, Multilayer Perceptron can be further used to diagnose the liver disorder efficiently. [12], demonstrated the predictive analysis of liver disorder using various classification algorithms. In this approach, Naïve Bayes and Support Vector Machine classification algorithms were used. These two algorithms were compared on the basis of performance parameters that include classification accuracy measures and execution time measures. The proposed system was implemented using Matlab 2013 tool and evaluated the dataset that had been collected from UCI Repository. After the experimental results, it had been observed that Support Vector Machine outperformed Naïve Bayes Algorithm due to the highest classification accuracy and can be used further in the prediction of liver disease

The early work by [4] employed simpler algorithms such as C4.5 and Naive Bayes, focusing primarily on their effectiveness for liver disease classification. However, [5] extended the comparison by including ten ML algorithms, with the addition of more complex techniques like SVM, KNN, and CHIRP. This shift indicates the growth of computational power and interest in assessing a broader range of algorithms to find the most suitable one for liver disease prediction.

In some studies, logistic regression and KNN performed well [4], whereas others showed a clear preference for more sophisticated techniques like Random Forest or SVM [8,5]. More recent studies demonstrate a preference for SVM and LR for their ability to handle complex, nonlinear relationships within medical datasets, hence SVM and LR was adopted for this research.

2.0. Methodology

Diagnosis of liver disease at a preliminary stage is important for better treatment. It is a very challenging task for medical researchers to predict the disease in the early stages owing to subtle symptoms. Various machine learning classification algorithms have been used in the prediction of liver diseases. In this study, we propose the use of Support Vector Machine (SVM) and Logistic Regression (LR) classification algorithm to identify the liver patients from healthy individuals.
The raw dataset for this study was collected from University of California Irvine (UCI), center for machine learning repository (http//archive.ice.edu/ml/datasets/). The liver datasets consists of 579 real world cases made up of 10 attributes having 1 categorical, 3 integers and 8 real numbers With 1 class and saved with CSV extension format in Microsoft 2010 version. The categorical attributes in the dataset were replaced with 0 and 1; the next stage was the normalization of the dataset. Normalization is the scaling down, transformation of attributes. The dataset were scaled to a range of (0,1) using min- max normalization equation [13]. The two major reasons for choosing the method was for Range Control: Min-max scaling transforms features into a common range, typically [0, 1] or [-1, 1]. This is beneficial for SVM and Logistic Regression since it prevents features with larger ranges from disproportionately influencing the model and Preserves Feature Relationships: Unlike standardization (which changes the mean and standard deviation), min-max scaling preserves the original relationships and distribution within the data. This approach is especially valuable if your dataset already has a uniform or bounded distribution. The min – max equation is represented in equation 1 below. Since both SVM and Logistic Regression depend on iterative optimization, scaling the data can lead to faster convergence by reducing the overall search space and improved accuracy: With normalized data, both models can learn the true patterns in the data more effectively, rather than being misled by features on different scales. This can lead to improved accuracy and a more stable decision boundary [14].

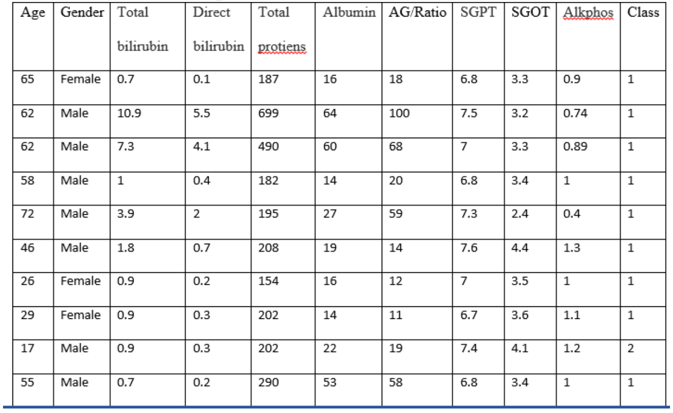
$$x_{ni} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Many factors affect the success of Machine Learning (ML) on a given task. If there are so much irrelevant (noisy), unreliable data and redundant information, then knowledge discovery during the training phase will be difficult. These can be addressed by data preprocessing. Data pre- processing includes data cleaning, normalization and transformation. The product of data pre- processing is the final training set. Even if you have a good data you need to make sure that it is properly scaled and in a format with meaningful features that best exposes the structure of the problem to the machine learning algorithm [15, 16].

3.0 Result and Discussion

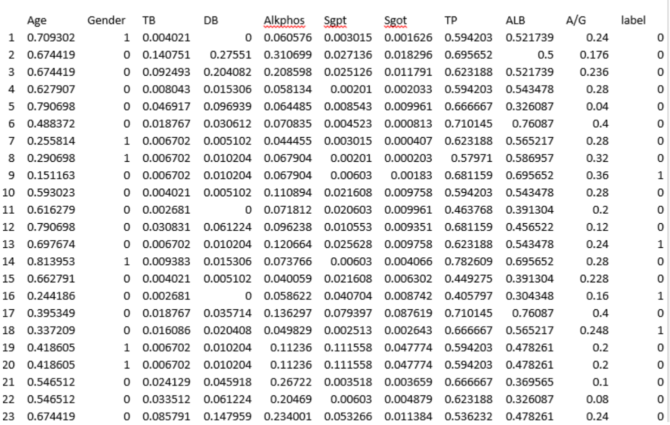
The raw dataset were saved with an extension of CSV format in Microsoft excel 2010 version. The preprocessing process started by cleaning the raw data, the raw dataset consists of categorical values real numbers and integers. A sample of the raw dataset and the scaled data set are represented in table 1 and table 2 respectively. The proposed system were trained with 405 dataset and tested with 174 dataset, the scaled dataset of 405 were inputted into the system and the classification algorithm was chosen. The models converged at various point after several iterations. The models' convergence rates were 90 epochs for LR and 4750 epochs for SVM the faster convergence of Logistic Regression in this liver disease prediction task is due to its simpler optimization landscape, efficient gradient-based optimization, and less dependency on specific points (like support vectors). For applications where quick convergence and high accuracy are essential, Logistic Regression is often a preferred choice due to its efficiency and lower computational burden [17] compared to SVM with the minimum error as captured in Figure 1 and Figure 2 respectively.

Table 1: Sample of raw liver dataset before scaling



Age	Gender	Total bilirubin	Direct bilirubin	Total protiens	Albumin	AG/Ratio	SGPT	SGOT	Alkphos	Class
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1

Table 2: Sample of the scaled liver dataset



	Age	Gender	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G	label
1	0.709302	1	0.004021	0	0.060576	0.003015	0.001626	0.594203	0.521739	0.24	0
2	0.674419	0	0.140751	0.27551	0.310699	0.027136	0.018296	0.695652	0.5	0.176	0
3	0.674419	0	0.092493	0.204082	0.208598	0.025126	0.011791	0.623188	0.521739	0.236	0
4	0.627907	0	0.008043	0.015306	0.058134	0.00201	0.002033	0.594203	0.543478	0.28	0
5	0.790698	0	0.046917	0.096939	0.064485	0.008543	0.009961	0.666667	0.326087	0.04	0
6	0.488372	0	0.018767	0.030612	0.070835	0.004523	0.000813	0.710145	0.76087	0.4	0
7	0.255814	1	0.006702	0.005102	0.044455	0.003015	0.000407	0.623188	0.565217	0.28	0
8	0.290698	1	0.006702	0.010204	0.067904	0.00201	0.000203	0.57971	0.586957	0.32	0
9	0.151163	0	0.006702	0.010204	0.067904	0.00603	0.00183	0.681159	0.695652	0.36	1
10	0.593023	0	0.004021	0.005102	0.110894	0.021608	0.009758	0.594203	0.543478	0.28	0
11	0.616279	0	0.002681	0	0.071812	0.020603	0.009961	0.463768	0.391304	0.2	0
12	0.790698	0	0.030831	0.061224	0.096238	0.010553	0.009351	0.681159	0.456522	0.12	0
13	0.697674	0	0.006702	0.010204	0.120664	0.025628	0.009758	0.623188	0.543478	0.24	1
14	0.813953	1	0.009383	0.015306	0.073766	0.00603	0.004066	0.782609	0.695652	0.28	0
15	0.662791	0	0.004021	0.005102	0.040059	0.021608	0.006302	0.449275	0.391304	0.228	0
16	0.244186	0	0.002681	0	0.058622	0.040704	0.008742	0.405797	0.304348	0.16	1
17	0.395349	0	0.018767	0.035714	0.136297	0.079397	0.087619	0.710145	0.76087	0.4	0
18	0.337209	0	0.016086	0.020408	0.049829	0.002513	0.002643	0.666667	0.565217	0.248	1
19	0.418605	1	0.006702	0.010204	0.11236	0.111558	0.047774	0.594203	0.478261	0.2	0
20	0.418605	1	0.006702	0.010204	0.11236	0.111558	0.047774	0.594203	0.478261	0.2	0
21	0.546512	0	0.024129	0.045918	0.26722	0.003518	0.003659	0.666667	0.369565	0.1	0
22	0.546512	0	0.033512	0.061224	0.20469	0.00603	0.004879	0.623188	0.326087	0.08	0
23	0.674419	0	0.085791	0.147959	0.234001	0.053266	0.011384	0.536232	0.478261	0.24	0

Table 2 showed the sample data raw set, the categorical value is column 2 (Gender), value one (1) was used to denote male while zero (0) for female. The dataset was spitted into 70% for training because training on more data reduces bias while testing on less data evaluates model generalization.

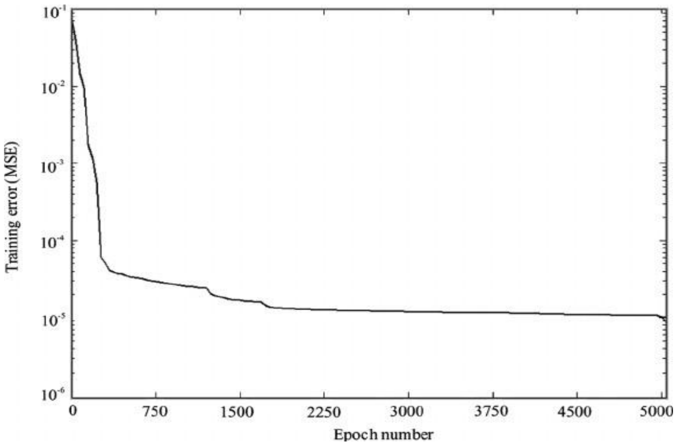


Figure 1: Convergence behavior of SVM

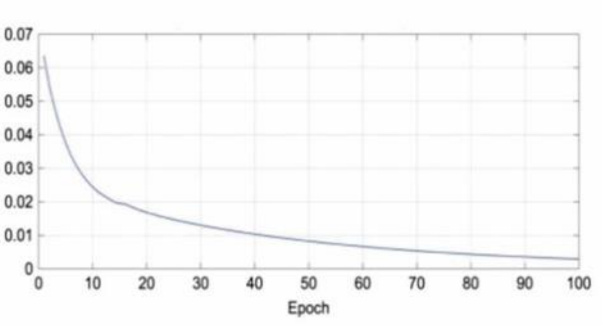


Figure 2: convergence behavior of Logistic Regression

In this study, we utilized some factual estimation that measures the test execution of the two classification algorithms. The performance of the classification methods was assessed by various evaluation procedures, such as accuracy, Precision and Recall. Consequently, the exhibition evaluation variables are determined by the confusion matrix. Here, True Positive (TP): The result of prediction correctly identifies that a patient has liver disease. False Positive (FP): The result of prediction incorrectly identifies that a patient has liver disease. True Negative (TN): The result of prediction correctly rejects that a patient has liver disease. False Negative (FN): The result of prediction incorrectly rejects that a patient has liver disease. Precision gives the contrast between sound and patient capacity ratio utilizing the prediction model. To discover the precision of classification is determined by the true positive, true negative, false positive and false negative.

Performance Measurement was carried out using SVM and Logistic Regression algorithm, the metric values obtained for the mentioned performance parameters by applying SVM Algorithm on the testing dataset are shown in the confusion matrix in Figure 3 while the confusion matrix for Logistic Regression algorithm is captured in Figure 4.

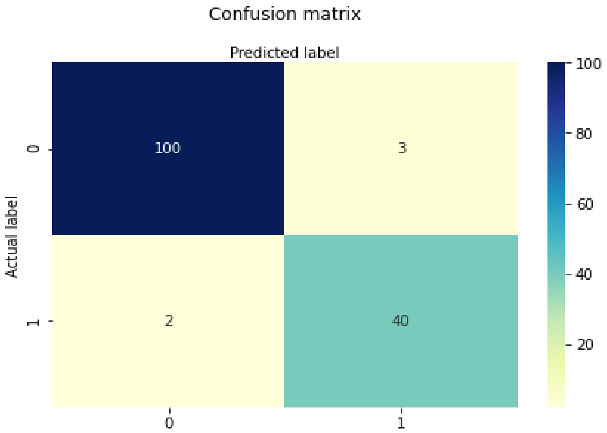


Figure 3: Confusion matrix on SVM classification

Where True positive is 100, False Positive is 3, False Negative is 2 and True Negative 40 with the following performance values; Accuracy of 96.55%, Precision value 0.97 and Recall: 0.98

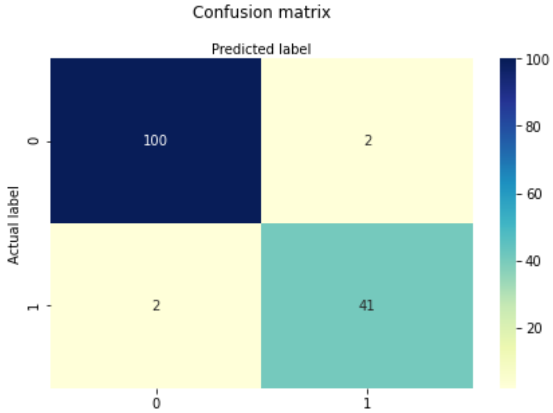


Figure 4: Confusion matrix on Logistic Regression classification

Where True positive = 100, False Positive = 2, False Negative = 2 and True Negative = 41 with Performance Accuracy of 97.24%, Precision: 0.98 and Recall: 0.98

4.0. Conclusion

Chronic Liver Disease is the leading cause of global death that impacts 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 liver function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy and many more. This disease diagnosis is very costly and complicated.

Liver disease counts to be one of the most prevalent diseases worldwide, thus this research employed two different classification algorithms namely Support Vector Machine and Logistic Regression for liver disease prediction. The comparison of all these algorithms were done based on classification accuracy, precision and recall which was found through confusion matrix. From implementation, Logistic Regression has the highest accuracy and precision of 97.24 % and 98% respectively. While SVM has classification accuracy of 95.55% and both algorithms have a recall of 98%. Confusion matrix gave the true positive and false positive rates of the classifier algorithms. Therefore, it can be concluded that Logistic Regression is more appropriate for predicting liver disease.

The demonstrated effectiveness of Logistic Regression with 97.24% accuracy and 98% precision suggests that it could be integrated into diagnostic systems to support clinical assessments. This could allow practitioners to make faster and more reliable diagnoses, potentially at a lower cost than traditional methods. Future research could explore more complex machine learning algorithms, such as neural networks or ensemble methods, which may offer improved accuracy, especially when combined with additional patient data (e.g., genetic information, lifestyle factors).

Reference

[1] Blachier MN, Leleu H, Peck-Radosavljevic M, Valla DC, RoudotHoraval F (2013), Burden of Liver Disease in Europe: A review of available epidemiological data. European Association for the Study of the Liver, Journal of Hepatology Vol. 58, Issue.3, pp 593-608

[2] Sindhuja D and Priyadarsini R. J. (2016), A survey on classification techniques in data mining for analyzing liver disease disorder", International Journal of Computer Science and Mobile Computing, Vol.5, no.5 pp. 483-488.

[3] [UK national statistics, http://www.statistics.gov.uk/ 2020

[4] Aneeshkumar A.S. and Venkateswaran C.J. (2012), Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Applications (0975 –8887) , Vol. 57, no. 6, pp. 39-42.

[5] Rashid N, Bilal K, Muhammad A, Karzan W, Atif K, Tapas R, Subhendu K. P (2016), Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset, International Conference on Computational Modeling and Security, India, pp. 862-870.

[6] Thirunavukkarasu K, Thirunavukkarasu K.K, Ajay Shanker S. G, (2018), Prediction of Liver Disease using Classification Algorithms 4th International Conference on Computing Communication and Automation (ICCCA)

[7] Kuppan P, and Manoharan N. (2017) "A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes", International Journal of Computing Algorithm, vol. 6, no. 1, pp. 2278-239.

[8] Gulia A, Vohra R, Rani P (2014), "Liver Patient Classification Using Intelligent Techniques," International Journal of Computer Science and Information Technologies (IJCSIT), vol. 5, no. 4, pp. 5110-5115.

[9] Ayesha P, Diksha M, Shrutika Jadhav , Rupali B, Rajeswari K (2018), Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver Disorder, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 6, Issue.2, pp. 388-394.

[10] Saritha B, Ramana S.V., Narra M, Rama P, Hiranmayi D, Eswaran K (2017), Classification of liver data using a new algorithm‖, 4th International Conference on New Frontiers of Engineering, Science, Management and Humanities, Hyderabad,

[11] Tapas R. B and Subhendu K. P (2016) "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security.

[12] Vijayarani S, and Dhayanand S (2015), Liver Disease Prediction using SVM and Naïve Bayes Algorithms‖, International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue.4 pp. 816-820.

[13] Dodge, Y. (2003). The oxford Dictionary of Statistical Terms OUP ISBN 0-19-920613-9

[14] Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University. Scaling the data can improve convergence and accuracy in both SVM and logistic regression models by reducing the search space and improving feature comparability.

[15] Teng, C.M. (1999). Correcting noisy data. In Proceedings of the Sixteenth International Conference on Machine Learning, 239-248.

[16] Liu, H. and Metoda, H. (2001). Instance Selection and Constructive Data Mining, The Kluwer International Series in Engineering and Computer Science 608 448 pp. I SBN 0 -7923 – 7209 -3

[17] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. - This work covers the computational efficiency of various optimization techniques and how simpler models like Logistic Regression often converge faster due to gradient-based methods.