IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LIVER DISEASE PREDICTION USING MACHINE LEARNING MODELS AND ALGORITHM

[1]N. Karthick, [2]S. Gowshik, [3]G. Saran, [4]C.S. Natesan, [5]A. Srinithi
[1]Assistant Professor, [2,3,4,5]BCA Final Year Students
[1,2,3,4,5]Dept. of Computer Applications, Sri Krishna Arts and Science College, Coimbatore, India.

Abstract: Liver disease has become a prominent global health concern, with conditions like cirrhosis andliver cancer ranking among the leading causes of mortality worldwide. The insidious nature of many liver diseases often results in asymptomatic onset, leading to underdiagnosis and delayed treatment. Early detection and management are pivotal in mitigating the impact of liver illnesses. Recognizing the challenges posed by the costly and intricate diagnostic processes, this study aimed to evaluate the effectiveness of diverse machine learning techniques in identifying liver disease. UT disease. Utilizing liver patient record dataset, this research explored five machine learning models for predicting the occurrence of liver disease using patients' medical histories. Through meticulous data pre-processing and analysis, encompassing tasks such as handling Managing absent data, encoding categorical variables, and standardizing features., the dataset was meticulously prepared for model training. Subsequentaluation of the Five algorithms for machinelearning, restated. revealed Random Forest as the top-performing model, achieving an accuracy of 75.87% on the test dataset. This research underscores the promise of machine learning in accurately predicting liver disease, thereby facilitating early diagnosis and intervention.

Keywords: machine learning, dataset, handling missing values and encoding categorical variables, Reformulated, normalizing features

1. Introduction

worldwide. Factors such as air pollution, poor dietary habits, excessive alcohol consumption, and misuse of medications contribute to the rising prevalence of liver diseases each year. This spectrum of conditions, including cirrhosis, liver cancer, and fatty liver disease, poses significant threats to personal health and overall well-being. Alarmingly, liver diseases rank among the leading causes of death globally. Of particular concern is the escalating incidence of non-alcoholic fatty liver disease, which now surpasses alcoholic liver disease in prevalence, with an annual growth rate of 25% [1]. Early diagnosis and Precise forecasting of liverstatus. diseases are paramount for effective treatment and improved patient outcomes. The Swift identification and prognosis of liver ailments play a crucial role in shaping treatment trajectories, outcomes, and the overall economic burden of healthcare [2]. Often, symptoms of liver disease remain subtle until the condition advances to severe stages, making early detection challenging. However, diagnosing liver disease in its initial phases enables prompt intervention and the implementation of tailored treatment plans, effectively averting further complications and potentially reversing the disease progression. Moreover, precise Forecasting liver illness essential for refining Segmenting patients and tailoring care delivery, and risk mitigation strategies. Utilizing prognostic models that leverage clinical data, genetic markers, and advanced computational techniques, healthcare providers can forecast disease progression and tailor treat approaches, aches accordingly, leading

to enhanced patient outcomes. Machine learning methodologies are emerging as a promising tool in healthcare, leveraging data analytics to bolster disease detection and prognostication capabilities. Machine learning has emerged as a critical asset in both healthcare research and clinical applications. By scrutinizing vast quantities Using machine learning models for medical data analysis. discern intricate patterns, unearth correlations, and produce forecasts grounded in observed trends. Across various healthcare domains, including diagnosis, treatment refinement, disease prognosis, and patient surveillance [3], machine learning techniques have proven invaluable. In clinical settings, machine learning algorithms empower clinicians to pinpoint abnormalities, such as tumors, fractures, or lesions, with a level of precision that may surpass human expertise. When it comes to Liver disease analyzed through machine learning algorithms utilize patient data to construct predictive models, aiding in the early detection and prognosis of the condition. The aim of this study was to examine a machine learning-based model capable of accurately identifying patients afflicted with liver disease. Utilizing a dataset sourced from Kaggle comprising Indian liver patient records, the study seeks to explore the effectiveness of five distinct Machine learning techniques such as Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting, reformulated. and K-Nearest Neighbors. Through thorough data analysis and comprehensive evaluation of the forecasting accuracy of these models, the study aims to uncover potential factors influencing their accuracy. Additionally, the research intends to compare its findings with existing literature in the field.

1.1 Literature Review:

Learned about the current state of nonalcoholic fatty liver disease [1]. Information on early disease detection and screening examination scheduling [2].Provides overview of machine learning applications in healthcare [3].

Studied how Logistic Regression applied to predict liver disease [4].

Reviewed Utilization of machine learning algorithms for predicting fatty liver [5].Studied the information regarding of support vector Machine [6].

Studied a powerful ensemble learning technique that has been widely applied in liver diseaseprediction, due to its ability to handle complex interactions and high-dimensional data [7].

Reviewed the application of the K-Nearest Neighbors (KNN) classification algorithm for Classification of large-scale medical health data, which could potentially be utilized for liver diseaseprediction by leveraging relevant medical data [8]. Studied the Forecasting liver disease utilizing gradient boosting machine learning methodologies, highlighting the importance of feature scaling to enhance the accuracy and effectiveness of the predictive models [9].

2. Data and Variables

The dataset containing records of liver patients in India, available on Kaggle, offers Comprehensive medical histories of patients diagnosed with liver disorders. This dataset comprises 583 samples, each representing the medical history of an individual patient, with 11 indicators provided for each sample. These indicators are outlined in Table 1. A value of 1 in the Dataset column denotes the presence of liver illness in the patient, while a value of 2 indicates otherwise. This dataset forms the cornerstone of our research project.Table 2 presents the average, variability, and maximum value, and minimum values of the numerical indicators in the dataset.

Table 1: description of the dataset used .

	count	mean	std	min	max
Age	583.00	44.75	16.19	4.00	90.00
Total_ Bilirubin	583.00	3.30	6.21	0.40	75.00
Direct_ Bilirubin	583.00	1.49	2.81	0.10	19.70
Alkaline_P hosphotase	583.00	290.58	242.94	63.00	2110.00
Alamine_A minotran-sferase	583.00	80.71	182.62	10.00	2000.00
Aspartate_ Aminotran-sferase	583.00	109.91	288.92	10.00	4929.00
Total_ Protiens	583.00	6.48	1.09	2.70	9.60
Albumin	583.00	3.14	0.80	0.90	5.50
Albumin_ and_Glo-bulin_Ratio	579.00	0.95	0.32	0.30	2.80

In the preliminary data analysis stage, the study examined within the breakdown between cases of liver disease and those without the dataset. It was found that 71.5% of individuals in the dataset had liver disease, while 28.5% did not. Moreover, this investigation delved into the proportion of genders and age distribution characteristics patient cohort to glean insights into the population under study. Statistical analysis revealed a notable disparity, with a significantly higher proportion of male patients compared to female patients. Additionally, a substantial proportion of male patients aged 30-65 was observed. Furthermore, data visualization unveiled that the age the spread of patients with liver disease irrespective of gender, generally followed a normal distribution. The depiction of gender and age distribution among liver disease patients and the general patient population is depicted in Figure 1.

Table 2: the evaluation metrics for every numeric attribute within the dataset.

Name	**Description**
Age	The patient's age
Gender	The patient's gender
Total_Bilirubin	The total amount of two classes of bilirubin, direct and Indirect bilirubin
Direct_Bilirubin	The amountof Direct Bilirubin
Alkaline_Phosphotase	The amountof AlkalinePhosphotase
Alamine_Aminotransferase	The amount of Alamine Aminotransferase
Aspartate_Aminotransferase	The amount of AspartateAminotransferase
Total_Protiens	The total amount of two classes of proteins, albumin And globulin
Albumin	The amountof Albumin
Albumin_and_Globulin_Ratio	Albumintoglobulin proportion
Dataset	Field that indicated whether a patient had liver disease or not

2.1 Cell Line and Reagents

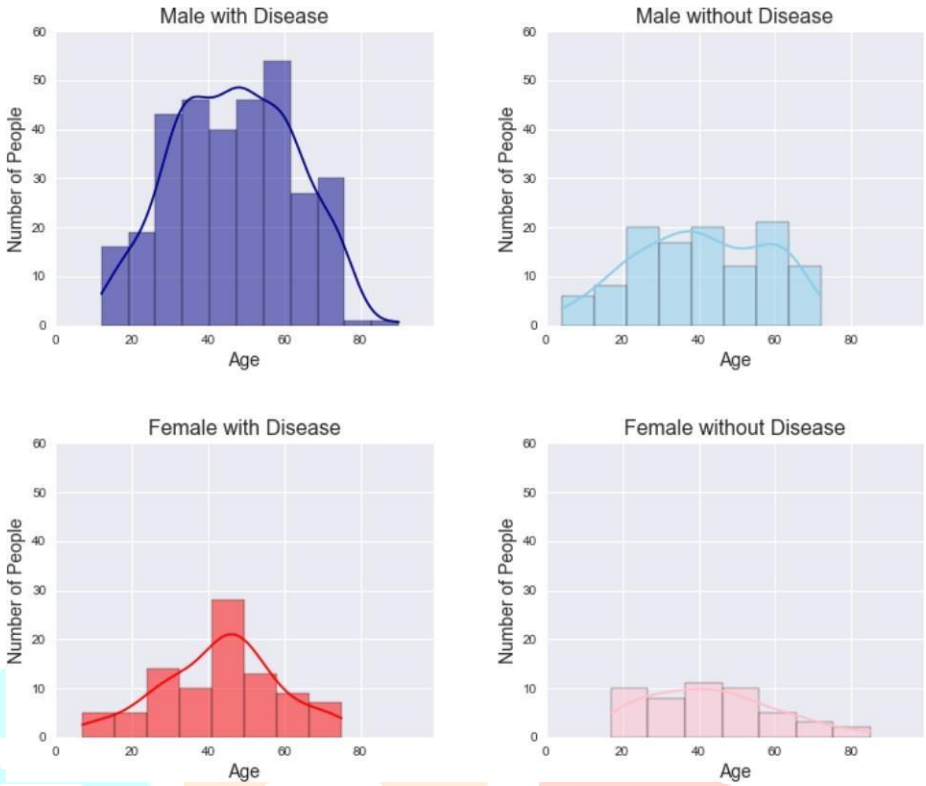

Fig 1: The arrangement of patients diagnosed with liver disease based on age and gender.

2.2. Method

Utilizing historical patient data, machine learning models can be trained to discern patterns and make predictions. Five commonly employed machine learning algorithms, namely Logistic Regression, Support Vector Machines, Random Forests, K-Nearest Neighbors, and Gradient Boosting, were employed to construct predictive models for detecting liver disease. Each algorithm possesses distinct characteristics which are is suited for diverse types of datasets and classification tasks. The study aims to investigate which model can attain superior prediction performance on the dataset.

2.2.1 Logistic Regression

Logistic regression presents an effective strategy for binary classification tasks. Through the logistic function, it models the relationship between the dependent variable (diagnosis of liver disease) and independent factors (indicators) [4]. This methodology offers valuable insights into the relevance andsignificance of individual features in the prediction process. The objective of logistic regression is todetermine the optimal parameters that maximize the likelihood of the observed output, a task achieved through Maximum Likelihood Estimation (MLE).

Linear Regression Formula:

$$f(x) = b + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n = w^T x$$

The formula for implementing Logistic Regression in Linear Regression:

$$h(x) = \frac{1}{1+e^{-f(x)}} = \frac{1}{1+e^{-w^T x}}$$

One of these functions, denoted as h(x), serves as an activation function known as the sigmoid function, illustrated in Figure 3. Resembling the shape of an "S," this function maps results to values between 0 and 1 [5].

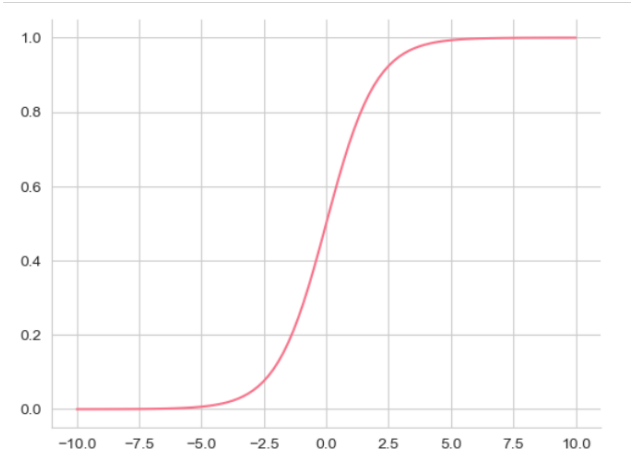

Fig 3: Sigmoid Function.

2.2.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) stand out as a powerful method capable of addressing both linear and nonlinear classification challenges. The primary objective of SVM is to identify the optimal hyperplane that effectively separates the dataset into different classes. This hyperplane serves as a decision boundary, distinguishing data points of one class from those of another. In SVM, the best hyperplane is determined by maximizing the margin, which refers to the sum of distances from the hyperplane to the closest data points of each class. To achieve this, the distances d1 and d2 from the samples on either side of the hyperplane are computed. The hyperplane with the largest margin, represented by margin = d1 + d2, is deemed the most suitable. SVM exhibits robustness in handling high-dimensional data and can capture intricate decision boundaries [6]. It's noteworthy that SVMs extend beyond binary classification tasks and can also be applied to multi-class classification and regression problems.

2.2.3 Random Forest

Both classification and regression tasks can be effectively addressed using Random Forest. This approach leverages ensemble learning, wherein multiple weak models are combined to form a robust model. In the context of Random Forest, these weak models are decision trees. To enhance accuracy and reliability, Random Forest constructs numerous decision trees and aggregates their predictions. The Random Forest algorithm creates forests by randomly selecting subsets of features for each tree split and training each tree on a distinct subset of the data [7]. By doing so, it mitigates the limitations of individual decision trees' weak generalization ability. While a single decision tree can only produce a single classification outcome, Random Forest aggregates the results of multiple trees, generating multiple classification outcomes. The final classification result is determined by the majority vote among these outcomes. Known for their resilience, scalability, and capacity to handle high-dimensional data, Random Forests are a preferred choice for various machine learning tasks.

2.2.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) stands as a versatile machine learning technique suitable for various applications, including regression, classification, and other tasks, although it is predominantly utilized for classification purposes. The algorithm operates on the fundamental principle of similarity, where similar items are proximate to each other. KNN functions by identifying the K nearest training samples (designated as K) and making predictions based on these samples [8]. This approach involves determining the distance between each sample and the unknown sample. Subsequently, the K known samples that exhibit the highest similarity to the unknown sample are selected. Under the majority voting rule, the category with the most votes among these K known samples is assigned as the category of the unknown sample. The straightforwardness and intuitive nature of KNN render it a favored choice for classification tasks.

2.1.2 Correlation analysis

In this investigation, associations between different characteristics and the presence of liver disease were examined, revealing pivotal variables with the strongest positive and negative correlations. Notably, thestudy uncovered that Albumin and Albumin_and_Globulin_Ratio exhibited the highest positive correlation coefficient at 0.16, while Direct Bilirubin displayed the largest negative correlation coefficient at -0.25. These findings underscore the significant influence of these indicators on the diagnosis and treatment of liverdisease in patients, emphasizing their importance in the diagnostic and therapeutic processes.

Professionals can utilize the magnitude of these indicators to devise more effective treatment strategies. To visually represent these associations, the authors generated a correlation coefficient heat map, as depicted in Figure 2.

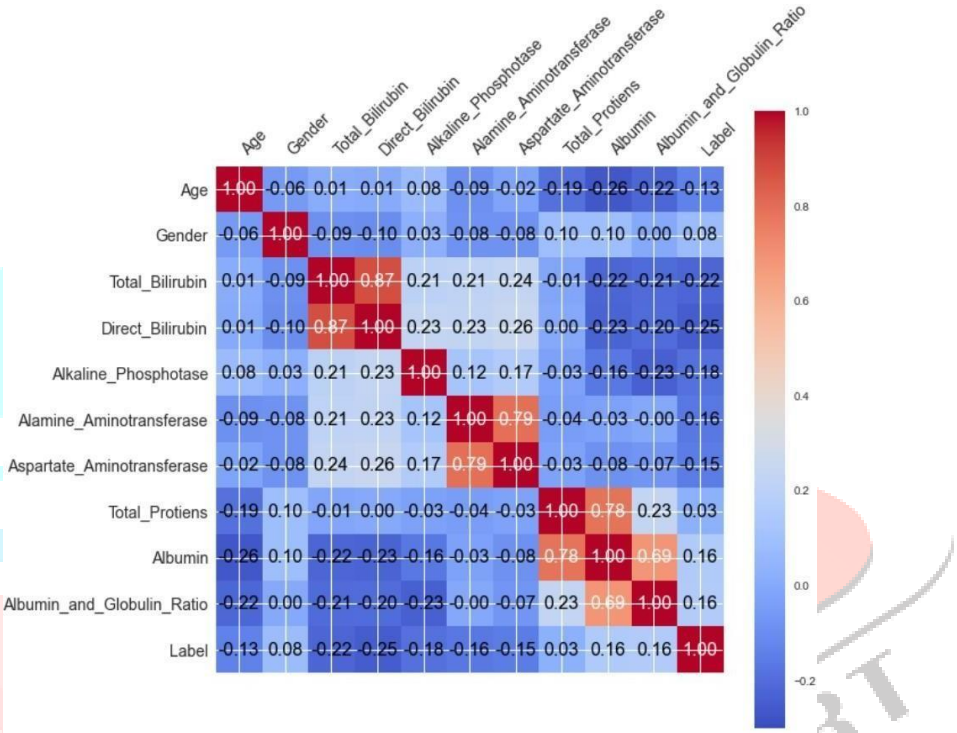

Fig 2: The heat map depicting the correlation coefficients of each feature in the dataset.

2.2.5 Gradient Boosting

Gradient Boosting, a powerful ensemble machine learning technique, is widely recognized for its ability to tackle both classification and regression challenges [9]. Belonging to a broader class of boosting methods, it focuses on minimizing the errors of preceding models by assigning greater weights to misclassified instances. Through iterative learning, Gradient Boosting progressively enhances these models, ultimately constructing robust predictive models. Notably, Gradient Boosting excels in handling complex interactions among predictor variables and has demonstrated remarkable success across diverse domains. The algorithm operates by building models sequentially, with each subsequent model trained to predict the residuals or errors of preceding models. These models are then aggregated to generate the final predictions. Gradient Boosting is commonly employed as a base model, particularly in conjunction with decision trees, including shallow trees.

3. Results and Discussion

For training the machine learning model, this research utilized a training set extracted from a dataset comprising Indian liver patient records. These models were trained on this dataset, acquiring insights into patterns and correlations between input features and the occurrence of liver disease.

To prepare the data for model training, the author standardized the features to ensure equitable comparisons and mitigate any potential biases stemming from differing scales. Furthermore, the dataset was partitioned into a training set (70%) and a testing set (30%), enabling the author to evaluate the performance of various machine learning models.

3.1 Performance Evaluation Indicators

The evaluation of model performance involved several key metrics [11]:

- Accuracy: This metric gauge the overall correctness of the model's predictions, representing the proportion of accurately categorized cases relative to all instances.
- Precision: It measures the percentage of accurately predicted positive occurrences (i.e., liver disease) among all anticipated positive instances, assessing the model's ability to minimize false positive outcomes.
- Recall (Sensitivity): This statistic indicates the percentage of positive examples that were correctly predicted out of all the positive instances that actually occurred, reflecting the model's capability to identify true positives while avoiding false negatives.
- F1-score: This metric represents the harmonic mean of recall and precision, providing a balanced assessment of the model's performance by considering both recall and accuracy.

3.2 Results of Performance evaluation on training set

The outcomes of the above-mentioned trained models are shown in Table 3.

Table 3: evaluation results of the test set for five distinct machine learning models.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.74	0.81	0.91	0.87
Support Vector Machines	0.73	0.75	1.00	0.84
Random Forest	0.76	0.86	0.87	0.89
k-Nearest neighbors	0.72	0.76	0.94	0.84
Gradient Boosting	0.71	0.78	0.87	0.80

From these findings, it is evident that the random forest model attains the highest accuracy of 0.82 onthe test set. Furthermore, it demonstrates robust performance across F1-score, recall, and precision metrics. These results highlight the superior predictive capability of random forest models in identifying the presence of liver disease with greater accuracy.

3.3 Results of Performance evaluation on training set:

The author compared the prediction scores of the five machine learning models between the training set and the test set. The results are presented in Table 4, with a graphical representation provided in Figure 4 to illustrate the comparison.

Table 4: scores of model predictions on the training and test sets.

Model	Train score	Test score
Logistic Regression	74.20	73.24
Support Vector Machines	100.00	74.69
Random Forest	80.31	70.41
k-Nearest neighbors	71.64	71.37
Gradient Boosting	95.81	68.27

Fig 4: comparison of model scores between the training set and test set.

3.4 Discussion

In research conducted by Ashwani Kumar and Neelam Sahu, they attained an accuracy of 79.22% employing the Random Forest algorithm, utilizing an 80-20% data split and six features [10]. In contrast, The Random Forest model utilized in this research endeavor attained an accuracy of 75.87% on the test set. The results from Kumar and Sahu's study indicate a higher the accuracy of the Random Forest model is relative to our findings. This variance I Variations in performance could be attributed to various factors, such as disparities in the dataset composition, variances in data pre-processing techniques, and variations in feature selection. It's noteworthy that the composition of the datasets utilized in the two studies might vary, leading to discrepancies in the data distribution and characteristics. Moreover, discrepancies in data pre-processing stages, such as handling missing values, scaling features, and encoding categorical variables, could impactthe efficacy of the models. Furthermore, the selection of features plays a pivotal role in the predictive capacity of machine learning models. Kumar and Sahu specifically utilized six distinct features [10]. The choice of features can significantly influence the model's ability to identify pertinent correlations and patterns within the data.

4. Conclusion

This study endeavors to leverage a dataset comprising Utilizing Indian liver patient records toconstruct machine learning models for liver disease. The project encompasses several Phases, encompassing dataset overview and pre-processing, Data analysis, machine learning model selection, training, and assessment stages.

Throughout The inquiry, the examination scrutinized the dataset, analyzed the spread of liver disease instances, and investigated the association of different attributes with the occurrence of liver disease. Additionally, data pre-processing steps were conducted to address missing values, standardize features, rename columns, and encode categorical variables. To facilitate model training and assessment, the dataset was partitioned partitioned into training and testing datasets

In selecting the appropriate machine learning model, we evaluated five popular algorithms: Logistic Regression, Random Forest, Support Vector Machines, Gradient Boosting, and K-Nearest Neighbors.Among these models, Random Forest emerged as the top-performing model, exhibiting the highest accuracy on the test set. This paper conducted a comparative analysis of the models using accuracy, precision, recall, and F1-score metrics. Despite achieving a test set accuracy of 75.87% with our random forest model, this result diverges from a related study by Kumar and Sahu [10], which obtained 79.22% accuracy using a similar approach. Variations observed can be attributed to disparities in datasets, data pre-processing techniques, feature selection, and other experimental factors. Further investigation is necessary to comprehensively comprehend these discrepancies and pinpoint potential areas for enhancement. In summary, this study underscores the utility of machine learning techniques in inferring liver disease from patient medical records. The results highlight the promise of machine learning models, particularly random forests, in accurately categorizing liver disease cases. Additionally, the findings underscore the importance of meticulous data pre-processing, Refinement of features and the process of choosing models enhance prediction accuracy. Future

research avenues may involve Investigating alternative approaches for selecting features, delving into Investigating alternative approaches for selecting features integrating additional pertinent clinical data to augment the forecasting accuracy of models. By continuallyrefining and enhancing Models for predicting liver disease, this contributes to timely identification and intervention, ultimately Enhancing patient healthcare results.

5. Reference

[1] T.G. Cotter, M. Rinella, Nonalcoholic fatty liver disease 2020: the state of the disease, Gastroenterology, 158, 1851-1864 (2020).

[2] S. Lee, H. Huang, M. Zelen, Early detection of disease and scheduling of screening examinations, Statistical Methods in Medical Research, 13, 443-456 (2004).

[3] S. Samarpita and R. N. Satpathy, Applications of Machine Learning in Healthcare: An Overview,2022 1st ICIDeA, Bhubaneswar, India, 51-56 (2022). https://doi.org/10.1109/ICIDeA53933.2022.9970177

[4] K. Sellamuthu, S. P, P. K and R. S, Liver Disease Prediction using Logistic Regression, 2022 8thICSSS, Chennai, India, 01-06 (2022). https://doi.org/10.1109/ICSSS54381.2022.9782179

[5] C.C. Wu, W.C. Yeh, W.D. Hsu. et al. Prediction of fatty liver disease using machine learningalgorithms, Computer Methods and Programs in Biomedicine, 170, 23-29 (2019). https://doi.org/10.1016/j.cmpb.2018.12.032

[6] W. Noble, what is a support vector machine. Nat Biotechnol, 24, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565

[7] L. Breiman, Random Forests, Machine Learning, 45, 5–32 (2001). https://dl.acm.org/doi/10.1023/A%3A1010933404324

[8] W. Xing and Y. Bei, Medical Health Big Data Classification Based on KNN ClassificationAlgorithm, in IEEE Access, 8, 28808-28819 (2020). http://dx.doi.org/10.1109/ACCESS.2019.2955754

[9] G. Shobana and K. Umamaheswari, Prediction of Liver Disease using Gradient Boost MachineLearning Techniques with Feature Scaling, 2021 5th ICCMC, Erode, India, 1223-1229 (2021). https://doi.org/10.1109/ICCMC51019.2021.9418333

[10] https://www.ijraset.com/fileserve.php?FID=7787

[11] https://www.analyticsvidhya.com