Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, No.3, February : 2025

MACHINE LEARNING MODELS BASED ON PREDICTIVE ANALYTICS FOR THE DIAGNOSIS OF LIVER DISEASES

Mr. Lalit Chaudhary, Assistant Professor, Computer Science, Himalayan Institute of Technology
Mr. Nitin Sharma, Ms. Garima Rathi, Assistant Professor, Uttaranchal School of Computing Sciences Uttaranchal University

ABSTRACT

Liver diseases affect about a million deaths worldwide. Liver problems can be diagnosed using several conventional techniques, but they are costly. All people at risk for liver disease might benefit from early liver disease prediction by receiving early. Machine learning greatly impacts healthcare as technology advances since it can predict illnesses early on. This study determines the predictive accuracy of machine learning for liver illness. The liver disease prediction (LDP) approach is presented in this article and can be used by researchers, stakeholders, students, and health professionals to forecast liver illness. Five techniques are available: K-Nearest Neighbours (K-NN), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA), but this research paper used Classification machine learning modelslike: Logistic Regression (LR), Support vector machine (SVM), Decision Tree (DT) and random forest (RF).To predict the result Python is comparing the accuracy. According to the results, the random forest algorithm providesthe best accuracy in predicting liver diseases.Over the allowable accuracy threshold and may be considered for the prognosis of liver disease.

Keywords: Liver, Machine learning, LDP, SVM, DT,LDA, Predictive

I. INTRODUCTION

Several important organs, each with a very beneficial function, are found in the human body. The liver is a solid large organ located below the ribs on the right-hand side of the body. It is located below the diaphragm, above the stomach, and right kidney.

The liver performs many functions. It plays a vital role in removing toxins, converting digested food to energy, storing vitamins and minerals, and controlling how much fat and sugar is sent back to the rest of the body.[1]

If the liver is ignored, it can be impacted by severaldiseases.[2]. The identification of liver-based diseases, which are among the world's top causes of death, is one of the most important areas of healthcare to benefit from the developments in machine learning and its integration with health data science.[3] When liver illnesses such as cirrhosis, fatty liver, hepatitis, and liver cancer are discovered in their advanced stages, there are few available treatments. Early detection lowers medical expenses while also greatly improving patient outcomes.

This synopsis focuses on developing and implementing machine learning algorithms to predict liver-based diseases. The aim is to leverage large datasets of medical records, and laboratory tests to create a predictive model that can identify individuals at risk of liver diseases.[4]. It is possible to enhance diagnostic accuracy and improve overall patient care.

In the changing atmosphere of health care and information technology, there's an adding occasion for the use of data wisdom and technology to epitomize health care and ameliorate the delivery of patient care. At its core, machine literacy (ML) utilizes artificial intelligence to induce prophetic models efficiently and more effectively than conventional styles through the discovery of retired patterns within large data sets. With this in mind, there are several areas within hepatology where these styles can be applied. [5]

Primary Liver Diseases

- **Fatty Liver** A reversible disease known as fatty liver occurs when big cholesterol fat vacuoles form in liver cells as a result of limit-setting. It might happen to those who have a high amount of alcohol use as well as to those who have never consumed alcohol.

- **Cirrhosis** One of the most dangerous liver conditions is. It is a procedure used to identify all types of liver disorders distinguished by the notable cell loss. The liver gets hard and leathery as it steadily shrinks in size. Under liver cirrhosis, the regeneration process persists, but the progressive loss of liver cells outweighs cell replacement.
- **Hepatitis** is typically brought on by a virus that is transferred by excessive contamination or close contact with bodily fluids that are infected.

II.LITERATURE REVIEW

Infectious complications that arise following liver transplantation (LT) are quite prevalent and contribute to increased mortality rates and prolonged hospital stays. Analyzing the impact of bilirubin and INR levels on the 5th postoperative day after pediatric LT through a two-factor binary regression model may facilitate early infection diagnosis and lessen the chances of adverse outcomes. This research indicates that the occurrence of infectious complications in the early phase after LT can be predicted with significant accuracy based on laboratory and calculated values of bilirubin and INR on the 5th postoperative day. Infections can deteriorate liver function, leading to higher levels of total bilirubin and INR.[6]

In this study, we present a novel technique for grading liver fibrosis utilizing fMRI hemodynamic response maps in response to hypercapnia . A Leave-One-Out evaluation of our dataset shows nearly 100% accuracy in distinguishing between no-fibrosis and low-grade fibrosis subjects, with an overall accuracy of 84.2% for grading fibrosis. These findings suggest that our method can, for the first time, accurately differentiate between healthy individuals and those with low-grade fibrosis using MRI images, achieving results comparable to traditional histology-based human grading. This technique could serve as a non-invasive alternative for classifying liver fibrosis patients and monitoring disease progression, as opposed to more invasive options like biopsy or contrast-enhanced imaging. We are in the process of expanding our database with an additional animal model for further investigation.[7]

Liver diseases are becoming more common over time, making it challenging to detect these conditions early on. Researchers have implemented numerous data mining models and machine learning techniques to identify such diseases in their early stages. However, in this area of liver disease prediction, it has been observed through experimental results that CHIRP is effective in reducing the error rate in evaluation metrics compared to other models used. When comparing performance, RF and MLP exhibit better accuracy than CHIRP. Nevertheless, the differences in accuracy between RF, MLP, and CHIRP are not significant when contrasted with the higher error rates mentioned.[8]

The suggested convolutional neural network was developed and executed. It was designed for the classification of uninfected liver images and images of metastasized (infected) lesions using TensorFlow. For image sizes of 65×65, 60×60, and 55×55, the highest classification accuracy achieved was 99% for the 65×65 image size. The proposed network was assessed against previous research and showed a significant improvement in classification accuracy. This enhancement was achieved through the implementation of a regularization technique to reduce the overfitting issue. Observations indicated that the F1 score is nearly one, demonstrating a balance between recall and precision. Therefore, it can be concluded that the proposed CNN model is the most effective for the binary classification of infected and uninfected liver CT images.[9]

As liver disease is challenging to diagnose due to the subtlety of its symptoms, this study is crucial for identifying the algorithms that exhibit the highest accuracy in forecasting this serious condition.Subsequently, five distinct supervised learning techniques are implemented using R (i.e., SVM, Naïve Bayes, K-NN, LDA, and CART), and the accuracy is measured using confusion matrix metrics. The findings indicate that K-NN achieves the highest accuracy at 91.7% for predicting liver disease. Autoencoders demonstrate slightly greater performance than K-NN because of their superior capability to identify overlapping features compared to traditional K-NNs. Most of the algorithms surpass the acceptable accuracy threshold of 75%.[10]

Liver Cirrhosis, which can be a life-threatening condition, demands urgent care to avert serious health complications. The implementation of machine learning models could greatly improve the early detection of cirrhosis, potentially minimizing its long-term harmful effects on health. A range of machine learning algorithms has been evaluated for their capacity to forecast liver infections based on various physiological markers, showing potential for future advancements in medical systems. These advancements are anticipated to enhance the precision and effectiveness of these tools. Furthermore, machine learning solutions could aid the public in evaluating the risk of severe conditions such as stroke in adults. Ideally, individuals with liver disease (LD) would gain from early identification and treatment, giving them a better opportunity to manage and recuperate from their illness.[11]

Identifying learning disabilities (LD) early is a costly and intricate task. To achieve this, we utilize the proposed model “back propagation (BK) artificial neural network (ANN) with incremental neurons in the hidden layer (HL).” We examined the performance of BK-ANN with hidden layer neurons ranging from 5 to 30 and compared it to various machine learning (ML) models. The ML models assessed include Decision Trees (DT), Naive Bayes (NB), Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Logistic Regression (LR). DT and NB achieved accuracies of 97.12% and 88.76%, ranking first and second, respectively. The training performance of the BK-ANN progressively improved as the number of hidden layer neurons increased. Ultimately, the BK-ANN with thirty hidden layer neurons exhibited exceptional performance compared to other BK-ANN configurations and ML models, achieving an accuracy of 99.9%.[12]

Description of the dataset

To solve the challenge of this paper, databases containing 583 records/entries are extracted from the Indian Liver Patient Dataset (ILPD) dataset. The UCI Machine Learning Repository . where this dataset was obtained. The complete ILPD dataset includes details on 583 liver patients from India. This includes 167 data not related to liver patients and 416 records related to liver patients. The northeast region of Andhra Pradesh, India, is where the data set was gathered. A selector is a class label that separates people into groups based on whether they have liver disease or not. There are following dataset details of attributes like age, gender, total bilirubin , alkaline phosphatase , alanine , aminotransferase and albumin.

Table: 2.1 Dataset Description

Name of Attributes	Data Type	Description
Gender	Boolean	Male=0 , Female=1
Age	Integer	Patient age in Year
Total Bilirubin	Float	Total bilirubin in the blood(mg/dl) High levels indicate liver problem
Direct Bilirubin	Float	Bilirubin level in the blood (mg/dL.) indicate bile flow
Alkaline Phosphatase (ALP**)**	Integer/float	Enzyme level in blood (U/L). High level indicate liver bile duct damage
Alanine Aminotransferase (ALT)	Integer/float	Liver enzyme level suggest liver cell damage
Aspartate Aminotransferase (AST)	Integer/float	It is also enzyme, indicates potential liver damage.
Total Proteins	Float	Amount of proteins in blood(g/dL).Low level indicate liver dysfunction
Albumin	Float	Show poor liver function
Albumin-Globulin Ratio	Float	Low ratio indicates liver damage
Liver Disease	Binary	1= liver disease present 0=No liver disease

III. Performance Evaluation Matrices

An essential component of any research project is model evaluation. You might get findings that meet your needs when you assess your model using a few common assessment measures. In this study, the suggested model is estimated using the following metrics compared to other models.

3.1 Accuracy is the most commonly used evaluation metric in classification problem. it measures the correctly predicted cases out of the total cases .

Accuracy $=\frac{\text{True }\mathbf{Positive}(\text{TP})+\mathit{True\ Negative}\ (TN)}{\text{Total }\mathbf{Number}\text{ of Predictions}}$

3.2 Recall (True Positive rate)

$$\boldsymbol{Recall} = \frac{\text{True }\mathbf{Positive}(\text{TP})}{\text{True Positive (TP)} + \mathit{False\ Negative}\ (FN)}$$

It measures how well the model identifies actual positives

3.3 Precision:

$$\boldsymbol{Precision} = \frac{\text{True }\mathbf{Positive}(\text{TP})}{\text{True Positive (TP)} + \mathit{False\ Positive}(FP)}$$

3.4 F1-Score:

$$\boldsymbol{F1\ Score} = 2 * \frac{\text{Precision . Recall}}{\text{Precision} + \text{ Recall}}$$

Provide a balance between precision and Recall

3.5 Confusion Matrix:

The positive Rate (Recall/Sensitivity)

False positiveRate :

$$\boldsymbol{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

False NegativeRate :

$$\boldsymbol{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

IV.Machine Learning Algorithms

To predict the best prediction used supervised Learning Algorithm **Supervised Learning** Algorithms labelled training data to learn the mapping function that turns input variables (X) into the output variable (Y). In other words, it solves for f in the following equation:

$Y = f\ (X)$

This allows us to generate outputs accurately when new inputs are given. [7]

4.1 Classification is used to predict the outcome of a given sample when the output variable is in the form of categories. A classification model might look at the input data and try to predict labels like —having liver disease‖ or —not having liver disease.

This research paper uses a supervised classification model to predict liver diseases.

4.1.1 Logistic Regression: The output of a categorical dependent variable is predicted via logistic regression. As a result, the result needs to be a discrete or category value. Yes or No, 0 or 1, true or false, etc., can be used, but probabilistic values that fall between 0 and1 .

Logistic regression equation

$Y{=}a_0{+}a_1x_1{+}a_2x_2{+}a_3x_3{+}\ldots\ldots{+}a_n\,x_n$in logistic regression Y can be between 0 and 1 so that it divides by (1-y). $\frac{y}{1-y}$

4.1.2 Support Vector Machine (SVM): SVM has attracted a lot of interest and is being actively tested for use in a variety of fields. SVMs are primarily utilized for learning ranking, regression, and classification functions. SVMs aim to identify the location of decision boundaries, often referred to as the hyperplane, that result in the best possible separation of classes. They are based on statistical learning theory and the structural risk minimization principle. [13]

4.1.3 Decision Tree: A decision tree that creates a double tree for categorization challenges. This approach is important in classification issues. This method uses a tree to complete the classification process, which is also applicable to a single record in the dataset and the item being classified for that record. The J48 algorithm simulates the lost values throughout this process; for example, the value for that element can be predicted based on how closely the categorization value for various records is perceived. The fundamental concept is to divide the data into runs according to the quality standards for the item identified in the working test. It allows classification using decision trees or rubrics that are constructed from scratch as well.[8]

4.1.4 Random Forestis a supervised classification method that is regarded as one of the most sophisticated ensembles learning techniques available, and it is a highly flexible classifier. As implied by its name, this method creates a forest comprised of multiple trees. While a significant number of trees are involved, in RF, the greater quantity of trees in the forest contributes to improved accuracy[14].

V. Machine learning Model performance analysis based on given matrices

5.1 Comparisons between different Matrices

The following Table: 2 and Figure :1 describe the matrices' performance using Machine learning model logistic regression (LR)

Table:2 Logistic Regression model comparison

Logistic Regression	
Matrices	Performance Analysis %
Accuracy	81.2
Recall	83.3
Precision	79.5
F1 Score	81.3

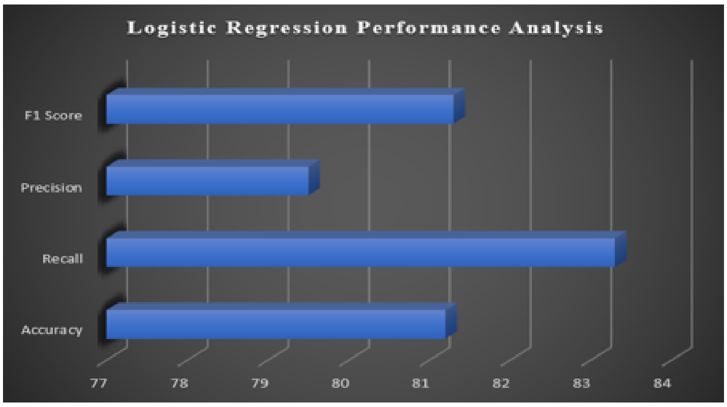


Figure: 1 Logistic Regression

The following Table: 3 and Figure:2 describe the matrices performance using support vector machine (SVM)

Table: 3 Support vector machine model comparison

Support Vector Machine (SVM)	
Matrices	Performance Analysis %
Accuracy	85.1
Recall	86.5
Precision	84
F1 Score	85.2

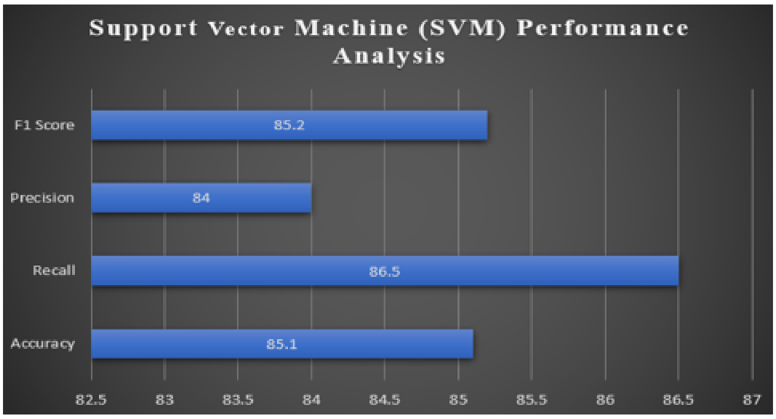


Figure: 2 Support vector Machine evaluation

The following Table: 4 and Figure: 3describe the matrices' performance using a Decision Tree (DT)

Table: 4 Decision Tree Model Comparison

Decision Tree (DT)	
Matrices	Performance Analysis
Accuracy	78.1
Recall	80
Precision	76.2
F1 Score	78

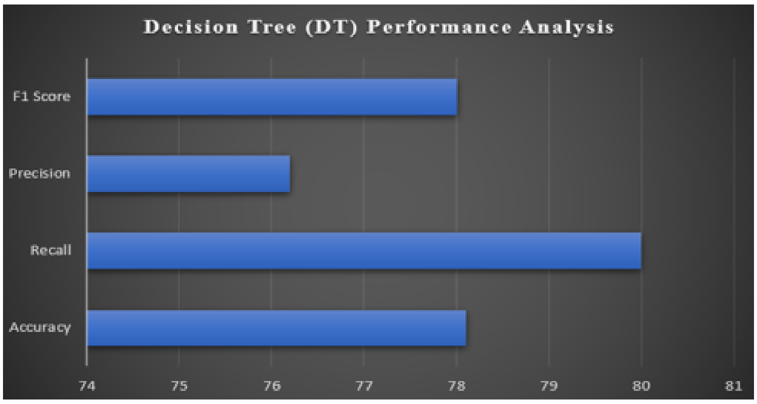


Figure:3 Decision Tree Evaluation

The following Table: 5 and Figure: 4describe matrices' performance using a Random Forest (RF)

Table: 5 Random Forest Model comparison

Random Forest (RF)	
Matrices	Performance Analysis
Accuracy	89.3
Recall	91.1
Precision	87.5
F1 Score	89.3

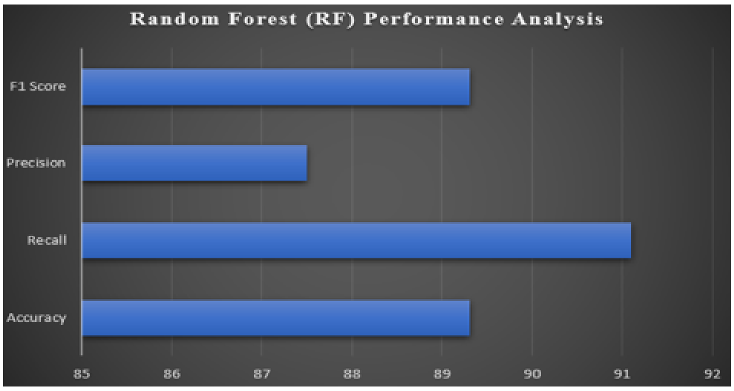


Figure:4 Random Forest (RF) Evaluation

VI. Confusion Matrix

The following Table and Figure: 5 evaluate the machine learning model using confusion matrix parameters TN, FP, FN, and TP.

Table: 6 ML model Evaluate using Confusion Matrix

ML Models	TN (True Negative)	FP (False Positive)	FN (False Negative)	TP (True Positive)
Logistic Regression	90	20	15	75
Decision Tree	85	25	20	70
Support Vector Machine	92	18	12	78
Random Forest	95	10	10	85

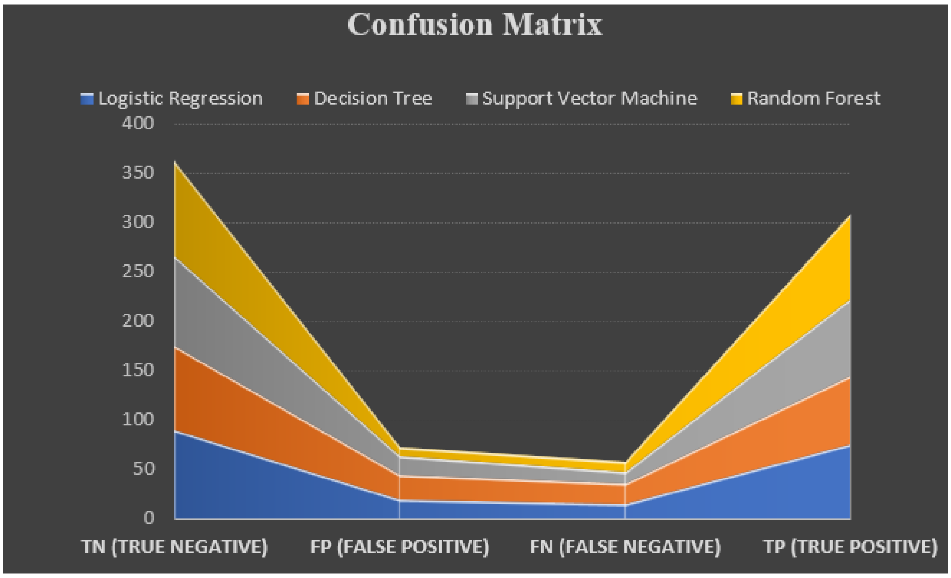
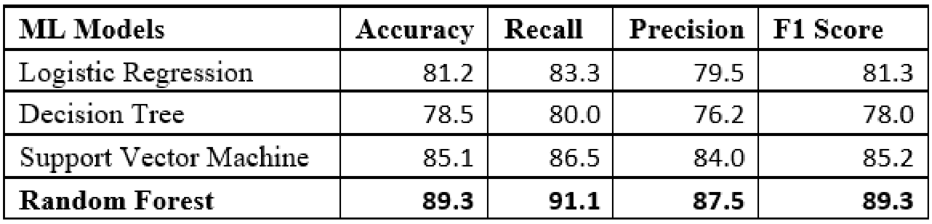


Figure:5 Confusion Matrix

VII.Analysis of Result

All models were evaluated based on the matrices test dataset. the result summarized compares all the graphs provided above.

Table: 7 Result Analysis



ML Models	Accuracy	Recall	Precision	F1 Score
Logistic Regression	81.2	83.3	79.5	81.3
Decision Tree	78.5	80.0	76.2	78.0
Support Vector Machine	85.1	86.5	84.0	85.2
Random Forest	**89.3**	**91.1**	**87.5**	**89.3**

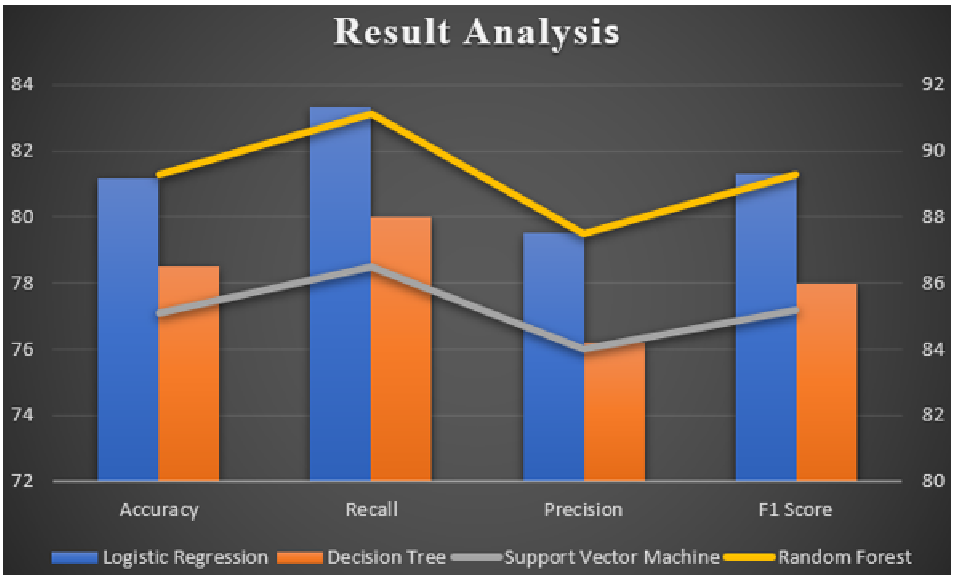


Figure:6 Result Analysis

VIII. Discussion

The Machine Learning algorithms for predicting liver diseases due to potential to enhance early diagnosis and improve patient outcomes. In this research paper, various studies have evaluated different machine learning algorithms to determine their effectiveness in this field. To predict the best result, compare different models like random forest, support vector machine, decision tree, and logistic regression.

IX. Conclusion

The main conclusion of this research paper, while using the machine learning models, mainly Random Forest, it considers dataset quality, interpretability, and clinical integration for their successful results in healthcare fields. Random forest shown high accuracy in identifying liver diseases. SVM and Decision Tree can be a secondary choice because of low performance with this dataset.

References

[1] G. Kadu, A. Engineer, S. Datta, and M. Polytechnic, “AN AUTOMATED LIVER DISEASE,” vol. 6, no. 2, pp. 25–28, 2018.

[2] A. Babic, U. Mathiesen, K. Hedin, G. Bodemar, and O. Wigertz, “Assessing an AI knowledge-base for asymptomatic liver diseases.,” *Proc. AMIA Symp.*, pp. 513–517, 1998.

[3] L. Philosophy and J. Jagtap, “DigitalCommons @ University of Nebraska - Lincoln

Bibliometric Review on Liver and Tumour Segmentation using Deep Learning Bibliometric Review on Liver and Tumour Segmentation using Deep Learning," 2021.

[4] J. Muthuswamy, "Extraction and classification of liver abnormality based on neutrosophic and SVM classifier," *Adv. Intell. Syst. Comput.*, vol. 713, pp. 269–279, 2019, doi: 10.1007/978-981-13-1708-8_25.

[5] T. A. Assegie, R. Subhashni, N. K. Kumar, J. P. Manivannan, P. Duraisamy, and M. F. Engidaye, "Random forest and support vector machine-based hybrid liver disease detection," *Bull. Electr. Eng. Informatics*, vol. 11, no. 3, pp. 1650–1656, 2022, doi: 10.11591/eei.v11i3.3787.

[6] I. Herliawan, M. Iqbal, W. Gata, A. Rifai, and J. J. Purnama, "Classification of Liver Disease By Applying Random Forest," *JTIK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 6, no. 1, pp. 89–94, 2020, doi: 10.33480/jitk.v6i1.1424.

[7] M. Banu Priya, P. Laura Juliet, and P. R. Tamilselvi, "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms," *Int. Res. J. Eng. Technol.*, vol. 5, no. 1, pp. 206–211, 2018, [Online]. Available: www.irjet.net

[8] B. Khan, R. Naseem, M. Ali, M. Arshad, and N. Jan, "Machine Learning Approaches for Liver Disease Diagnosing," *Int. J. Data Sci. Adv. Anal.*, vol. 1, no. 1, pp. 27–31, 2019, doi: 10.69511/ijdsaa.v1i1.71.

[9] E. Engineering, "Electronics Engineering, Department of Technology, Shivaji University,Kolhapur,Maharashtra,India Electronics & Telecommunication Department, KITS college of Engineering, Kolhapur, Maharashtra, India," vol. 4, no. 1, pp. 64–73, 2022.

[10] M. M. Bushra, "Liver Disease Detection using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 8, pp. 156–162, 2024, doi: 10.22214/ijraset.2024.63879.

[11] Fahmudur Rahman, Denesh Das, Anhar Sami, Priya Podder, and Daniel Lucky Michael, "Liver cirrhosis prediction using logistic regression, naïve bayes and KNN," *Int. J. Sci. Res. Arch.*, vol. 12, no. 1, pp. 2411–2420, 2024, doi: 10.30574/ijsra.2024.12.1.1030.

[12] P. Terlapu, R. P. Sadi, R. Pondreti, and C. Tippana, "Intelligent Identification of Liver Diseases Based on Incremental Hidden Layer Neurons ANN Model," *Int. J. Comput. Digit. Syst.*, vol. 11, no. 1, pp. 1027–1050, 2022, doi: 10.12785/ijcds/110183.

[13] D. M. Srivenkatesh*, "Performance Evolution of Different Machine Learning Algorithms for Prediction of Liver Disease," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 2, pp. 1115–1122, 2019, doi: 10.35940/ijitee.l3619.129219.

[14] F. Muhammad *et al.*, "Liver Ailment Prediction Using Random Forest Model," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1049–1067, 2023, doi: 10.32604/cmc.2023.032698.