Applying Machine Learning Algorithms for Liver Disease Prediction

K.V. Narasimha Reddy[1*], Satish Duggineni[2], Munaf Shaik[3], Anjibabu Bandaru[4], D. Venkata Reddy[5], Dr. Sireesha Moturi[6]

[1]Asst.Professor,Department of CSE, Narasaraopeta Engineering College, Narasaraopet-522601, Palnadu, Andhra Pradesh, India.
[2,3,4]Department of CSE, Narasaraopeta Engineering College, Narasaraopet-522601, Palnadu, Andhra Pradesh, India.
[5]Asst.Professor,Department of CSE, Narasaraopeta Engineering College, Narasaraopet-522601, Palnadu, Andhra Pradesh, India.
[6]Assoc.Professor,Department of CSE, Narasaraopeta Engineering College, Narasaraopet-522601, Palnadu, Andhra Pradesh, India.
narasimhareddynec03@gmail.com ;

Abstract. Liver disease poses a significant global health concern, particularly in countries like India. Early detection is crucial for effective treatment but remains challenging due to the delayed onset of symptoms. This study utilizes various machine learning algorithms to forecast liver disease based on patient data. The models used include Support Vector Machine (SVM), K-Neighbors, Hard Voting Classifier, Multilayer Perceptron, Decision Tree, Logistic Regression, Random Forest, and Genetic Algorithm optimization. Performance metrics such as Accuracy, Precision, Recall, and F1-Score were employed to assess model performance. The Random Forest model optimized with Genetic Algorithm achieved the highest accuracy of 79%, making it the most effective model for liver disease prediction. This approach aids in faster and more accurate diagnoses, enhancing clinical decision-making.

Keywords: Feedforward neural network · Perceptron method · SVM · K-Neighbors · Random forest classifier· Decision Tree · Voting-based classifier · Logistic Regression · Genetic Algorithm (GA).

1 INTRODUCTION

Liver disease is a growing global health concern, with significant mortality and morbidity rates, particularly in countries like India. Conditions like cirrhosis, hepatitis, and non alcoholic fatty liver disease are prevalent, driven by factors like viral infections, alcohol consumption, and lifestyle changes.
Despite the availability of advanced diagnostic tools, early detection remains challenging because symptoms often appear in the later stages when the disease has progressed [1]. This delay in identification raises the possibility of deadly consequences and complicates treatment.

The increasing concern on the worldwide burden of liver diseases naturally enhances interest to apply machine learning approaches in improving early detection. With a good ma chine learning model applied on patient data that includes liver function test, among other biochemical markers, it is possible to identify the disease patterns that are otherwise not easily detected by other methods [2]. This approach not only improves diagnostic accuracy but also allows for earlier intervention, which is crucial for reducing the severity of the disease.
The predictive models, such as SVM, Neural Networks, and Voting Classifiers have shown considerable success in liver disease prediction. These models can efficiently handle large datasets, process complex relationships, and offer insights that assist in clinical decision making [3]. As a result, these methods represent a significant advancement in healthcare technology, providing a more data-driven approach to diagnosing liver disorders.
Early-stage detection through these predictive models has proven beneficial in identifying high-risk patients even before clinical symptoms develop. This capability is vital in reducing liver disease-related mortality and morbidity, as it enables timely treatment and better disease management [4]. Machine learning integration in health care is not only helpful in the establishment of early diagnosis of liver conditions but also provides a scalable solution to this ever-growing burden related to liver disease.
Machine learning algorithms present a promising solution to the challenges of diagnosing liver disease. As these technologies advance, they enable healthcare professionals to achieve more accurate, efficient, and early detection methods. This progress represents a significant advancement in enhancing patient outcomes and mitigating the global impact of liver disease [5].
This paper proposes the application of multiple machine learning algorithms to predict liver diseases using patient data. The models used in this study include SVM, K-Neighbors, Hard Voting Classifier, Multilayer Perceptron, Decision Tree, Logistic Regression, and Random Forest, optimized with a Genetic Algorithm. Key performance metrics such as Accuracy, Precision, Recall, and F1-Score are used to evaluate the models.

2 LITERATURE STUDY

Research in the applicability of machine learning algorithms in diagnosing liver disease has gained significant momentum recently. Various studies have demonstrated the effectiveness of different methodologies, providing insights into improving diagnostic processes.

One such study by Murty and Kumar [6] used a multi-layer perceptron neural network to improve classifier accuracy in liver disease diagnosis. This research highlights how advancements in neural network architectures can significantly enhance the precision of liver disease classification, offering a promising approach to diagnostics.

Haque et al. [7] compared the performances of random forests and artificial neural networks in the classification of liver disorders. Random forests exhibited

robustness, while artificial neural networks excelled at identifying complex patterns. This comparative analysis is essential for selecting the right model based on the specific needs of a diagnosis.

Joloudari et al. [8] integrated Particle Swarm Optimization (PSO) with Support Vector Machines (SVM) and feature selection techniques in a computer-aided decision-making study. This research emphasizes the importance of optimization techniques to improve SVM performance, leading to better liver disease predictions through enhanced feature selection.

Gaber et al. [9] explored the use of supervised learning methods combined with genetic algorithms for the automatic classification of fatty liver disease. The findings indicate that genetic algorithms optimize the learning process, leading to better classification results for fatty liver disease, which is vital for effective patient management.

Kuzhippallil et al. [10] conducted a comparative analysis of machine learning algorithms tailored for Indian liver disease patients. This study highlights the variations in model performance across different patient demographics and emphasizes the need for customized approaches to ensure effectiveness in such a diverse population.

Gupta et al. [11] provided an overview of various machine learning classification algorithms for predicting liver disease. This comprehensive study contributes to a broader understanding of how machine learning can be utilized for accurate liver disease diagnosis.

Musleh et al. [12] demonstrated the potential of artificial neural networks in predicting liver disease with high accuracy. The research underscores the utility of neural networks in clinical settings for early diagnosis and intervention. Overall, these studies form a significant body of work that establishes the progress in machine learning techniques for liver disease diagnosis. Importantly, the studies have made insightful observations about how various models and optimization strategies might be further developed in this critical area of healthcare. The proposed study is analyzing the various liver function tests that are in employment in order to give a prediction regarding liver disease, considering a set of patient data as input and putting through several classifiers, namely: SVM, K-Neighbors, Voting Classifier, Multilayer Perceptron, Decision Tree, Logistic Regression and Random Forest. Performance metrics that are used are the ones that would give the best model on the basis of the predicted liver health: Precision, Recall, F1-score, Accuracy and Confusion-Matrix.

3 METHODOLOGY

The methodology in the below figure presents a typical workflow for machine learning. It begins with using the collected data and preprocessing followed by data cleaning and normalization to prepare it to feed into the system. Then, the dataset needs to be divided to be split for training and testing purposes. For the training of the model, a genetic algorithm was employed with the objective of feature selection while optimizing the input variables of the model being trained.

Lastly, the model was trained, and the performance of that model was estimated by evaluating accuracy results.

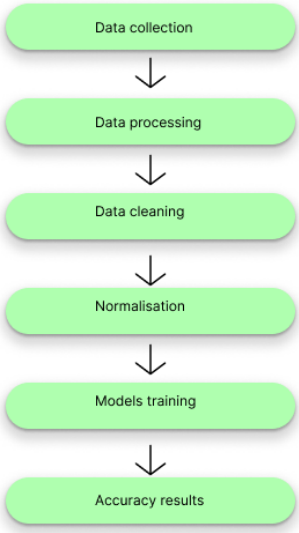

Fig. 1. Workflow Model.

3.1 DATA COLLECTION

This liver patient's dataset is based on the result of bio chemical tests, the demographic details of the patient, and the target label to indicate whether the patient has liver disease or not. It contains 583 records in 11 attributes [13].

- Age: Age of the patient in years (integer).This gives insight into how age is related to the disease in the liver.
- Gender: Gender of the patient in concern Male or Female.This attribute may be useful when analyzed for gender differences to observe the prevalence of the disease across genders.
- Total Bilirubin: total bilirubin level in milligrams per deciliter (Float).Based on the amount of total bilirubin in the blood, it shows the condition of the liver.
- Direct Bilirubin: Direct bilirubin level in mg/dL- float.Another very relevant parameter of the liver function is represented by the direct bilirubin. It is the level of bilirubin that has already been extracted and transported to the liver for further treatment.

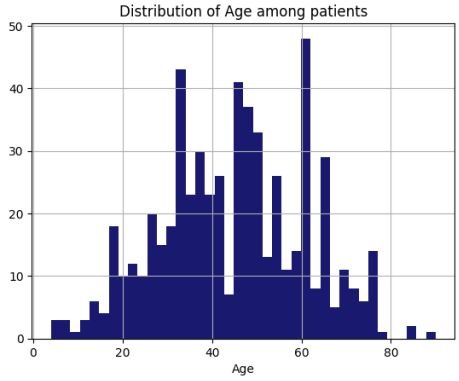

Fig. 2. Distribution of age among patients.

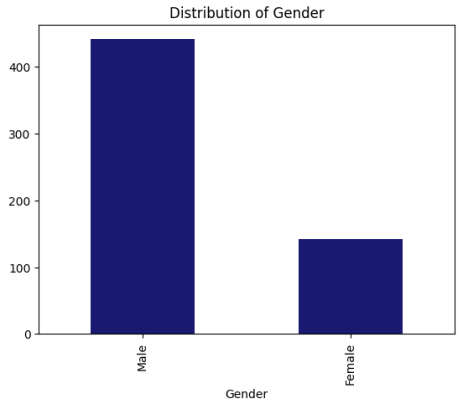

Fig. 3. Patients gender count.

- Alkaline Phosphotase: Levels of alkaline phosphatase enzyme (int).High levels can point to disorders of theliver or bones.
- AlamineAminotransferase: Levels of ALT enzyme(int).Increased levels over normal of ALT signal inflammation or injury in the liver.
- Aspartate Aminotransferase: Levels of AST enzyme (int).It helps in ascertaining liver damage, more in relation to the level of ALT.
- Total Protiens:Total protein level in gm/dL(float).This is a test meant for calculating the total amount of protein present in the blood, which in turn becomes an indicator of liver function.
- Albumin: Albumin is a protein level measured in grams per deciliter (gm/dL) (float).If albumin levels are low, this may indicate chronic liver disease.
- Albumin and Globulin Ratio: The ratio of albumin to globulin protein in blood (float). An abnormal ratio may indicate liver malfunction.

– Dataset: Classification label, in which 1 denotes the liver patient and 2 the non liver patient. It has been used as a target variable for machine learning tasks in order to predict liver disease.

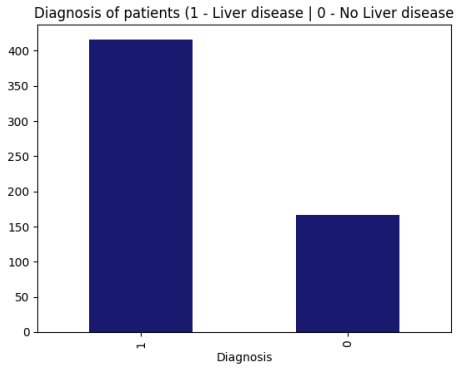

Fig. 4. Healthy and Unhealthy liver count.

3.2 DATA PROCESSING

Pre-processing is very much essential for the enhancement of the dataset with the addition of necessary features such as scaling of data, removal of irrelevant columns, and filling in gaps of missing values that were available from the original database. Immediately after loading the data, visualization of key patterns is an important aspect. Also, the format of data needs to be changed into a proper structure with good classification and visualization. This includes data standardization, feature consistency, and exhaustive cleaning of data in order to get rid of noise or inconsistencies from it and to make it ready for machine learning and further usable at analyses.

3.3 DATA CLEANING

The data have several impurities that need to be cleaned to improve the model accuracy prediction. The dataset may contain nominal and categorical features, such as gender, pre processed into a numerical variable. In addition, the dataset contains a number of missing fields and null values. KNN imputer imputation technique is used for filling in gaps , so that the model developed couldn't sacrifice accuracy or reliability.

3.4 NORMALIZATION

The Normalization technique scales each feature in such a way that it is very close to a standard normal distribution with mean 0 and standard deviation 1.

Since the range of features is quite different in the input dataset, standardization removes such differences and scale it to reduced values. Therefore, it makes the model building simpler and also helps in choosing an appropriate activation function for the perceptron algorithm. In the Indian Liver Patient Dataset, there was a target feature of a binary classification visualized in a pair plot. However, all the data points lay highly scattered and overlapped, hence not linearly separable. To predict correctly, several nonlinear separability-handling models were adopted.

3.5 MODELS

SUPPORT VECTOR MACHINE (SVM) SVM model utilizes the RBF kernel for non linear classification. The parameter 'C=100' is used in order to handle regularization, higher values of which would mean less regularization, and the fit closer to the training data. The parameter 'gamma=0.0001' defines how much of influence each training point would have, and with higher values, the decision boundaries would be smoother. All these are made towards achieving accuracy in classifications.

K-NEIGHBORS (KNN) The K- Neighbors model takes n-neighbors = 4, meaning for each data point it considers the classes of the four nearest neighbors to make a prediction. It calls the class voted by the majority of these neighbors. This is the parameter that controls how local or global the decision would be done; having fewer neighbors means the model is more sensitive to what is local [14].

MULTILAYER PERCEPTRON NEURAL NETWORK (MLP) MLP-Classifier model uses GridSearchCV to 'tune' some of its most important parameters: hidden-layer sizes=(11, 9, 1) gives the network structure, activation is set at tanh and ReLU and both SGD as well as Adam are taken as solver options. Learning-rate is constant or adaptive, max iter values 200 as well as 400 are applied. Cross-validation is applied to find the best parameters with the goal of getting the highest possible accuracy score for the model.

HARD VOTING CLASSIFIER (HVC) VotingClassifier with SVC, Decision Tree, and Logistic Regression. Logistic Regression uses default settings, while the Decision Tree uses gini for splits and best for the splitter. The VotingClassifier applies hard voting, predicting based on the majority class from all three models. This ensemble method promotes improvement in classification accuracy by combining the strengths of each model.

RANDOM FOREST (RF) Random Forest model uses a genetic algorithm to optimize three key hyperparameters: n estimators between 10 and 200 , max-depth between 10 and 200 , and min-samples-split also between 10 and 200. The

random-state is fixed at 21 for reproducibility. The genetic algorithm maximizes model accuracy as the fitness function, and the best hyperparameters are used to train the final model.

DECISION TREE (DT) This Decision Tree Classifier uses a genetic algorithm to select the key hyperparameters: max-depth is chosen between 2 and 20, min-samples-split is also between 2 and 20, and criterion must either be gini or entropy. The random-state was set to 21 so that the model could potentially be replicated. To optimize these parameters so the model hit peak levels of accuracy, model accuracy was used as the fitness function [15].

LOGISTIC REGREESION (LOGREG) Logistic Regres sion model uses a genetic algorithm that optimizes two hy perparameters: one such parameter is regularization strength, which ranges from 0.01 to 10.0, and the other is penalty, which could either be l2 or l1. Since the l1 solver has been assigned to liblinear, for l2 it must be lbfgs. The maximum iterations to stop the training have been set to 1000 along with a fixed random-state as 21 so that the results would be replicated. The fitness values used are the accuracies to decide the optimal set of hyperparameters.

GENETIC ALGORITHM (GA) Genetic algorithm is a optimization algorithms. It selects the features by genetically. Also it has parameters like initial population, mutation rate, crossover rate. It takes the features based on the best fitness score and accuracy can improve by using this optimisation algorithm.

3.6 RESULT AND ANALYSIS

The Indian liver patient data-set was assessed with a combination of machine learning models which include K- Neighbors, Hard Voting Classifiers, Multilayer Perceptron Neural Networks, and SVM. Optimization of more model types such as LogReg , RF, and DT by the help of Genetic Algorithms enhances their predictions. The various approaches therefore helped in making comparisons on the accuracy of different models in predicting the outcome of liver disease for this patient group.

Figures 1, 3 indicates a distribution of the patient by age groups, whether they bear a disease of the liver. For this study, 1 will be used to indicate that the patient's liver is diseased whereas 2 will be put in place when the patient's liver is healthy and has no trace of having a disease. The dataset consists of 583 individuals, with 71% ,413 people identified as having liver disease, while 28.64% ,170 individuals were reported with healthy livers. Graph showing that the greatest number of casualties of liver diseases falls within the age bracket of 40-60 years.

Figure 4 presents a pair plot, showcasing the relationships between each feature in the dataset. This visual representation offers key insights that help in determining the most appro- priate predictive algorithm, ensuring optimal performance for the dataset. As the problem states that classification is binary , the graph shows that the data points are very overlapped and scattered., making it impossible to separate them using a single linear decision boundary. As a result, a nonlinear model is necessary for effectively addressing this dataset and achieving better prediction accuracy.

Figure:5 illustrate the correlations between two variables can be measured us-

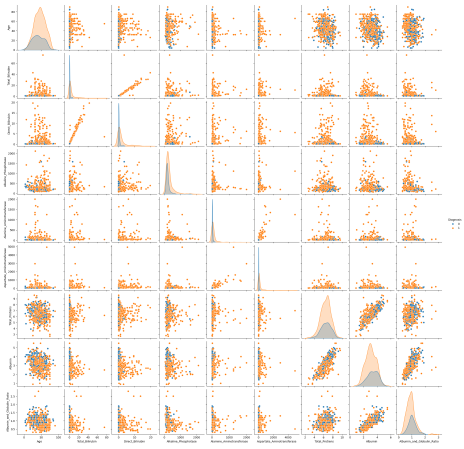

Fig. 5. Pair plot among the attributes.

ing correlation coefficient ranging from negative one (-1) meaning that the two variables have an inverse relationship to a positive one (+1) denoting that they possess a direct relationship while Zero (0) means no correlation at all. However, Figure:5 demonstrates that there is a clear correlation present within the dataset, with various features exhibiting different levels of relationship. This visual representation highlights how certain variables are correlated, positively aiding or negatively in understanding the underlying data structure.

Table 1 presents the optimization metrics for various models analyzed in liver disease dataset. Evaluating all indices, including accuracy, precision, recall, F1-score, and the confusion matrix, is crucial for determining the most suitable model for the dataset. Among the models, the Random Forest with Genetic Algorithm achieves the highest accuracy of 79%, with strong precision 81%, excellent recall 94%, and an F1 Score of 0.87, indicating a well-balanced performance.

As per table 2 it shows the analysis of the exsiting system which is taken from the reference paper [3]. The table cleary states and shows the accuarcy,f-score,Precision,Recall and specficitiy of the existing models.

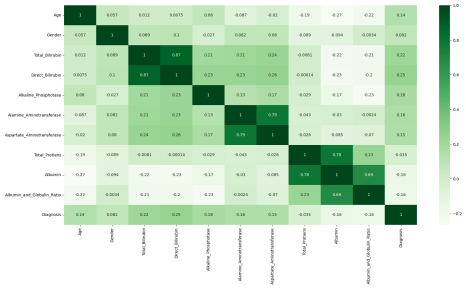

Fig. 6. Correlation matrix based on attributes of dataset.

Table 1. Performance Metrics of Models

Models	Accuracy	Precision	Recall	F1 Score
SVM	0.754	0.75	1.00	0.86
KNN	0.745	0.79	0.91	0.84
MLP	0.746	0.78	0.92	0.84
HVC	0.756	0.76	0.97	0.85
RF(GA)	0.79	0.81	0.94	0.87
DT(GA)	0.75	0.75	1.00	0.86
LOGREG(GA)	0.78	0.80	0.95	0.87

The comparison accuracy of the existing model got the highest accuracy of 78%, f-socre got 87% whihc is simillary to the f-score for the proposed model and the highest accuracy in proposed model is Random Forest Algorithm using gentic Algorithm which gives 79%. By this comparison we can cleary say that the propsed model got the highest accuracy .

Table 2. Analysis of the Existint System.

Models	**Accuracy**	**F-score**	**Precision**	**Recall**	**Specificity**
MLP (11, 9, 1)	0.7724	0.8715	0.7724	1	1
SVM (ker = RBF, c=100, g=0.0001)	0.7655	0.864	0.7826	0.9643	0.9643
KNN (K-NN, K=4)	0.7310	0.8368	0.7874	0.8929	0.8929
HVC (ker = RBF)	0.7862	0.8724	0.8092	0.9464	0.9464

Logistic Regression model, optimized with a Genetic Algorithm, achieves an accuracy of 78%, with an impressive 80% precision, a high recall rate of 95%, and an F1 Score of 0.87. The Support Vector Machine exhibits strong performance with 75.4% accuracy, a precision of 75%, perfect recall at 100%, and an F1 Score of 0.86, emphasizing its recall strength. K Nearest Neighbors secures a 74.5% accuracy, with 79% precision, 91% recall, and an F1 Score of 0.84, demonstrating robust recall but slightly lower accuracy. The Multilayer Perceptron mirrors this

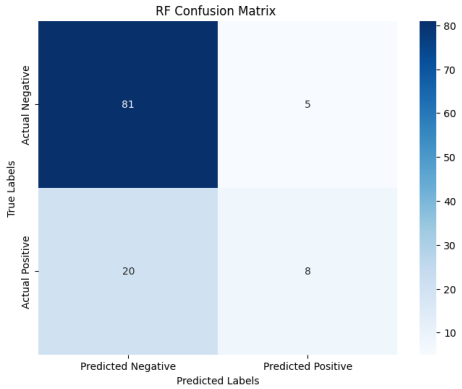

Fig. 7. Visual representation of Performance Metrics for each model.

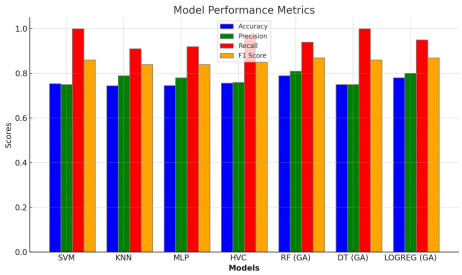

Fig. 8. Visual representation of Performance Metrics for each model.

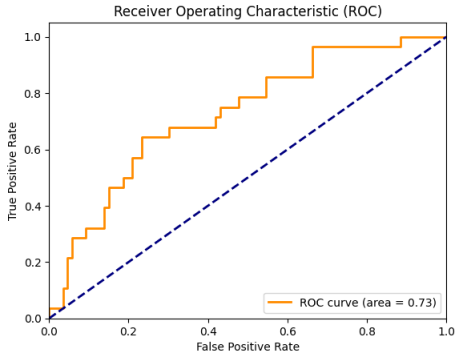

Fig. 9. ROC curve of random forest classifier.

with 74.6% accuracy, 78% precision, 92% recall, and an F1 Score of 0.84. The Hard Voting Classifier achieves 75.6% accuracy, 76% precision, 97% recall, and an F1 Score of 0.85, showcasing its recall efficiency. The Decision Tree, combined

with a Genetic Algorithm, delivers 75% accuracy, 75% precision, perfect recall of 100%, and an F1 Score of 0.86, indicating its capability inidentifying positive cases with balanced performance.

Figure 9 ROC curve illustrates the balance between the true positive rate (sensitivity) and the false positive rate for this classification model. With an area under the curve (AUC) of 0.73, the model demonstrates moderate performance in distinguishing between classes, outperforming random chance.

3.7 CONCLUSION

Liver disease prediction is an essential aspect of modern healthcare, aimed at identifying liver conditions at an early stage to prevent severe health consequences. The human liver acts as one of the most vital organs for detoxification, protein synthesis, and many functions associated with digestion, and it can come under several kinds of diseases like liver cancer, fatty liver disease, cirrhosis, and hepatitis.
Early detection is important, as liver diseases often progress without noticeable symptoms until they reach advanced stages, making treatment difficult. This delay in seeking medical attention worsens the condition, and liver disease is often diagnosed too late, making treatment difficult and increasing the risk of death. This study focuses on the potential of identifying liver disorders through liver function tests derived from blood work, enabling the detection of liver disease in its early stages, ensuring timely treatment and prevention of severe outcomes.

The results of the study suggest that after applying the various predictive models, including SVM, K-Neighbors, Multilayer Perceptron, Hard Voting Classifier, Logistic Regression with GA, Decision Tree with GA, and Random Forest with GA, each model was analyzed through key performance metrics including accuracy, precision, recall, and F1 Score. Among these models, Random Forest with Genetic Algorithm achieved the highest accuracy of 79%, with strong precision 81%, excellent recall 94%, and an F1 Score of 0.87. Therefore, we can conclude that, for this dataset, the Random Forest with Genetic Algorithm provides the best and at most highest accuracy. In future ,there is a chance to explore higher-order optimization techniques and deeper learning architectures, along with a larger and more diverse dataset, in order to further enhance the accuracy of predictive models for detection of liver disease.

References

1. Thuluvath, Paul J., Anoop Saraya, and Mohamed Rela. "An introduction to liver disease in India." *Clinical Liver Disease* 18, no. 3 (2021): 105-107.
2. Devarbhavi, Harshad, Sumeet K. Asrani, Juan Pablo Arab, Yvonne Ayerki Nartey, Elisa Pose, and Patrick S. Kamath. "Global burden of liver disease: 2023 update." *Journal of Hepatology* 79, no. 2 (2023): 516-537.

3. Anthonysamy, Victor, and SK Khadar Babu. "Multi Perceptron Neural Network and Voting Classifier for Liver Disease Dataset." *IEEE Access* (2023).
4. Dutta, Krittika, Satish Chandra, and Mahendra Kumar Gourisaria. "Early-Stage detection of liver disease through machine learning algorithms." In *Advances in Data and Information Sciences*, pp. 155-166. Springer, Singapore, 2022.
5. Moturi, Sireesha, Jhansi Vazram Bolla, M. Anusha, M. Mounika Naga Bhavani, Srikanth Vemuru, SN Tirumala Rao, and Sneha Ananya Mallipeddi. "Prediction of Liver Disease Using International Conference on Data Science and Applications." In *Data Science and Applications*, pp. 243-254. Singapore: Springer Nature Singapore, 2023.
6. Murty, Sivala Vishnu, and R. Kiran Kumar. "Enhanced classifier accuracy in liver disease diagnosis using a novel multi-layer feed-forward deep neural network." *International Journal of Recent Technology and Engineering* 8 (2019): 1392-1400.
7. Haque, Md Rezwanul, Md Milon Islam, Hasib Iqbal, Md Sumon Reza, and Md Kamrul Hasan. "Performance evaluation of random forests and artificial neural networks for the classification of liver disorder." In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1-5. IEEE, 2018.
8. Joloudari, Saadatfar, Javad Abdollah Hassannataj, Dehzangi, Hamid and Shahaboddin Shamshirband. "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection." *Informatics in Medicine Unlocked* 17 (2019): 100255.
9. Gaber, Ahmed, Hassan A. Youness, Alaa Hamdy, Hammam M. Abdelaal, and Ammar M. Hassan. "Automatic classification of fatty liver disease based on supervised learning and genetic algorithm." *Applied Sciences* 12, no. 1 (2022): 521.
10. Kuzhippallil, Maria Alex, Carolyn Joseph, and A. Kannan. "Comparative analysis of machine learning techniques for Indian liver disease patients." In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 778-782. IEEE, 2020.
11. Gupta, Ketan, Nasmin Jiwani, Neda Afreen, and D. Divyarani. "Liver disease prediction using machine learning classification techniques." In *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 221-226. IEEE, 2022.
12. Musleh, Musleh M., Eman Alajrami, Ahmed J. Khalil, Bassem S. Abu-Nasser, Alaa M. Barhoom, and SS Abu Naser. "Predicting liver patients using artificial neural network." *International Journal of Academic Information Systems Research (IJAISR)* 3, no. 10 (2019).
13. Indian liver patient dataset."https://www.kaggle.com/datasets/ jeevannagaraj/ indian-liver-patient-dataset"
14. Moturi, Sireesha, Jhansi Vazram Bolla, M. Anusha, M. Mounika Naga Bhavani, Srikanth Vemuru, SN Tirumala Rao, and Sneha Ananya Mallipeddi. "Prediction of Liver Disease Using Machine Learning Algorithms." In *International Conference on Data Science and Applications*, pp. 243-254. Singapore: Springer Nature Singapore, 2023.
15. Mamidala, Sai Kumar, Sireesha Moturi, SN Tirumala Rao, Jhansi Vazram Bolla, and KV Narasimha Reddy. "Machine Learning Models for Chronic Renal Disease Prediction." In *International Conference on Data Science and Applications*, pp. 173-182. Singapore: Springer Nature Singapore, 2023.