Design and Development of Prediction of Liver Disease, its Seriousness and Severity using Machine Learning

[1]Rajesh Prasad, [2]Achal R. Mate, [3]Pushkar A. Narkhede

[1,2,3]*MIT School of Engineering, MIT Art Design and Technology, Pune, India*

***Abstract*— At present time Liver Disease is the most dominant medical issue all over the world, as about 2 billion patients die. Due to the fact that liver illness is frequently discovered at a late stage, which results in mortality. Early Diagnosis and treating patients are compulsory to reduce the risk of that lethal disease, but no system detects liver disease in the early stage. In order to help the medical community begin treating liver damage properly, this article intends to create a system that can identify liver illness in its early stages as well as the severity of the condition. Additionally, it can reduce the chances of patient death and liver transplant and allow doctors to understand the severity of the liver. In order to classify the Indian Liver Patient Dataset (ILPD), we tested several machine learning (ML) techniques, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Ada Boost Classifier (ADB), and Xgboost Classifier (XGB). The experimental results showed that AdaBoost Classifier performs well with the highest accuracy of 99 percent. The data set has been investigated with regard to precision, TP rate, FP rate, and Kappa statistics. This classification system is evaluated for accuracy using a confusion matrix. After that Severity prediction is done using Rule Engine, It is fully customizable by Doctors on medical attributes to predict the severity of liver disease in different geographical locations, and Environment conditions that effects on medical attributes of liver disease.**

***Keywords*— Liver Disease, Mortality, Prediction, Severity, Hybrid Machine Learning, Classification, Clustering, Rule Engine.**

I. INTRODUCTION

Recent years have seen an increase in the number of researchers creating automated disease prediction models utilizing supervised machine learning techniques. The likelihood of patients dying from diseases can be reduced with an early liver disease diagnosis. Our project aims to develop a system that can detect the liver disease in the earlier stage as well as the severity of the detected liver disease so that the patient gets the early meditation and proper treatment to recover from liver disease. In this study, an effective automated liver disease prediction model is created using a machine learning algorithm approach with hybrid modeling and predicting the severity of that liver disease by using the medical history, characteristics, pathological test, and blood report information. To do this, we use a simple rule engine with specific medical characteristics that predict the seriousness of the liver disease and are integrated into a web app using a flask. In the proposed model, the data is entered by the user through and web app, the analysis is then performed in a real-time by using a pre-trained machine learning model which is already integrated into the web app, and finally, a liver disease detection along with its severity is shown on the web app. In our system, we are generating the Lab report where the report consisting of the patient detailed symptoms and the result of our ML model. Comparative investigation shows that the suggested model can assist doctors in prescribing treatments drugs in a timely manner.

II. LITERATURE SURVEY

In the past years, several types of research have been done on liver disease classification. Vasan D, Suyan R, Dinesh K "Liver Disease Prediction using machine learning" [1]: The medicinal enzyme features from the UCI repository dataset on ILPD are extracted in this research project using data mining techniques like PSO Feature Selection (Indian Liver Patient Dataset). Predicting liver disease is the major goal. The prediction process basically consists of five parts. To normalise the data, the Min-Max technique is first used to the ILPD. Significant features are then extracted utilising PSO feature selection techniques. The classification of liver diseases occurs in the third phase, which employs algorithms like J48, Naive Bayes, and Support Vector Machine (SVM). The best classifier with a high accuracy and precision score is evaluated using evaluation metrics in the fourth phase. A comparison of each classifier is made in the final stage for selecting outperforming classifiers.

Rakshith D B, Mrigank S, Ashwani K, Gururaj S P "Liver Disease Prediction System using Machine Learning techniques"[2]: The primary goal of the research is to estimate the likelihood of developing liver disease using the UCI ILPD's medical characteristics (Indian liver patient dataset). In this research work, a variety of machine learning classifiers is applied to UCI ILPD to predict the risk of liver disease using all features and on 582 instances. The classifier SVM (support vector machine) is 100% accurate when compared to KNN (K-Nearest Neighbors), ANN (artificial neural network), and Naive Bayes.

Md Fazle R, S M Mahedy H, Arifa I C, Md Asif Z, Md Kamrul H " Prediction of Liver Disorder using machine learning algorithm: The goal of this study is to forecast liver illness utilizing useful medical characteristics from the UCI ILPD (Indian Liver Patient Dataset). The comparative study of 5 different machine learning techniques is done to accurately classify if the patient is suffering from liver disease or not. Comparing the experimental result, the boosting classifier (AdaBoost) is found as the best classifier with PCC-FS (Pearson Coefficient Correlation - Feature Selection). The PCC-FS is applied on UCI ILPD to extract the significant features to evaluate the relationship between the dependent and independent features.

Sateesh A, Vijayalaxmi A, Rashmi U, Shruthi P, Vilaskumar P "Optimizing Liver Disease Prediction with random forest by various data balancing techniques" [4]: The main focus of this research is the optimization of Random forest algorithm using hyperparameter tunning, rather using different machine learning algorithm directly. Different data balancing techniques are applied to UCI ILPD to balance the data. Cluster Centroids, Condensed Nearest Neighbour, and ALLKNN data balancing techniques are used. Random forest (RF) with different statistical methods such as Feature selection, data transformation PCA (Principal Component Analysis), and data balancing experiment. For final experimentation confusion matrix and classification report with more focus on precision, recall, and ROC (Receiver Operating Characteristics) results were compared.

Javad H, Hamid S, Abdollah D, Shahaboddin S " Computer-aided decision making for predicting liver disease using PSO based optimized SVM with feature selection"[5]: Identification of significant features from a UCI ILPD is the purpose of this research work. For feature significance and importance tree-based modelling methods such as information gain, Entropy, and Gini impurity measure are used as PSO. Selected features are also subjected to a range of machine learning methods, and their effectiveness is assessed using the FPR (False Positive Rate), ROC (Receiver Operating Characteristics), Accuracy, and F measure.

S. Vijayarani, S. Dhayanand el at " Liver Disease prediction using SVM and Naïve Bayes Algorithm" [6]: The major goal of this study is to use UCI ILPD to predict several forms of liver illness, including cirrhosis, hepatitis, and liver cancer. The Nave Bayes algorithm and SVM (Support Vector Machine) are used to predict liver illness. The classification of liver disease type is carried out by taking into account the results of lab tests and liver function tests.

Kalyan N, Amulyashree S "NeuroSVM: a graphical user interface for Identification of Liver Patient" [7]: Data mining techniques are used in this study's classification of liver disease. On the UCI ILPD, the naive bayes, random forest, bagging, and ANN (artificial neural network) algorithms are used. Additionally, a hybrid neuroSVM model combining SVM and a feed-forward neural network was created to improve accuracy. The feed-forward neural network's input is the SVM prediction outcome based on actual data.

Meng Z, Changjun S, Tao L, Tianyue H, Shiming L "Fatty Liver Disease Prediction Model Based on Big Data or Electronics Physical Examination records" [8]: The researchers mainly focus on the prediction of the fatty liver based on health records collected from the hospital. The dataset is also subjected to the chi-squared test, with a significance level of 0.05 percent, to determine the significance of the features. The xgboostst (Extreme Gradient Boost) algorithm is then used to make the final prediction, with Bayesian optimization on the hyperparameter and triple cross-validation.

Jagdeep S, Sachin B, Ranjodh K "Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques"[9]: Worked on UCI ILPD dataset for identification of liver disease using software engineering approaches and machine learning techniques. Without feature selection, the efficiency of the machine learning algorithm is low was analyzed. In their proposed work feature selection techniques are used for comparison of different machine learning algorithms.

Ivair P, Denio D, Guilherme D B, Julyana F L "Exploratory Analysis of Electronic Health Records using Topic Modelling" [9]: The researchers mainly focus on Exploratory Data Analysis for features importance, a Significance level of Electronics Health RecordsThe topic modeling technique is used for information discovery and topic extraction for better insights from data. LDA (Linear Discriminant Analysis) is applied to a dataset for discovering the latent topics from the records.

Nazim Razali "A data mining approach to the prediction of liver disease"[10]: Researchers work on UCI ILPD to the extraction of important features which are useful for the classification of liver disease. Using a variety of hybrid machines and deep learning approaches the final classification is done. Additionally, for a highly accurate approach metrics such as precision, recall, and accuracy is analyzed. The author's main focus is only on the improvement of accuracy, precision, and recall metrics using different hybrid approaches and data mining techniques

Thirunavukkarasu K, Ajay S S, Md Irfan, Abhishek C el at "Prediction of liver disease using classification algorithm" [11]: The goal of the authors is to use a different machine learning algorithm to categorize whether a patient has liver disease or not. The performance metric is employed for the final classification and accurate approach, and sensitivity and specificity are given greater attention.

Fahad M, Easy H, Morgan W, Hafiz K, "Statical Machine Learning Approaches to Liver disease prediction" [12]: Using Machine learning and statistical techniques a classifier is applied to selected and significant data. Hypothesis testing is implemented for identifying significant features from medical records. PCA (Principal Component Analysis) technique is applied to data to reduce the dimension of the big dataset obtained from the hospital. To find the relevant PC, variable importance ranking using the Gini index is used (Principal Component). On the selected PC's variety of Machine learning algorithms are applied as a binary classifier for the prediction of liver disease.

Keerthana PSM, Nimish P, Riya M, Koppula Bhanu P R, Nidhi L "A Prediction Model of Detecting Liver disease in a patient using Logistic Regression of Machine Learning" [13]: Researchers significantly utilized the Logistics Regression as a pre-processing technique for data pre-processing. Additionally ANN (Artificial Neural Network), and RF (Random Forest) are utilized for accuracy in classification. Using ROC (Receiver Operating Characterises) the final result is drawn.

Insha A, Chiranjit D, Tanupriya C, Abha T "Liver Disease Detection due to excessive alcoholism using data mining techniques" [14]: Researchers want to employ data mining techniques to provide an early forecast of liver damage brought on by excessive alcohol use. By applying various data mining approaches such as SMO (Sequential Minimal Optimisation), Bayes Net, and J48 to the dataset the best approach is finalized considering the accuracy. Before classification data clustering is done on instances showing similar test results.

L. Alice Auxilia "Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease" [15]: The author's main focus is to predict the liver malady (hepatic infection) of liver disease. Additionally, common machine learning algorithms are utilized for achieving higher accuracy by considering the performance metrics. As a pre-processing step correlation and significant features are extracted. Comparison of algorithm done by using confusion metrics by focusing on the accuracy, precision, and recall score.

III. RESEARCH OBJECTIVE

The majority of the time, the early signs of liver illness are mild and difficult to spot. Even a healthy individual with fatty liver disease does not exhibit any symptoms in the early stages, and it eventually results in liver transplantation or death. The greatest solution is to predict liver disease at an early stage. This study aims to forecast the severity and onset of liver disease at an early stage. We can identify liver disease at an early stage by reviewing the results of the linked blood tests and liver function tests, and by taking into account the typical ranges of the enzymes, we can classify the expected liver disease's severity.

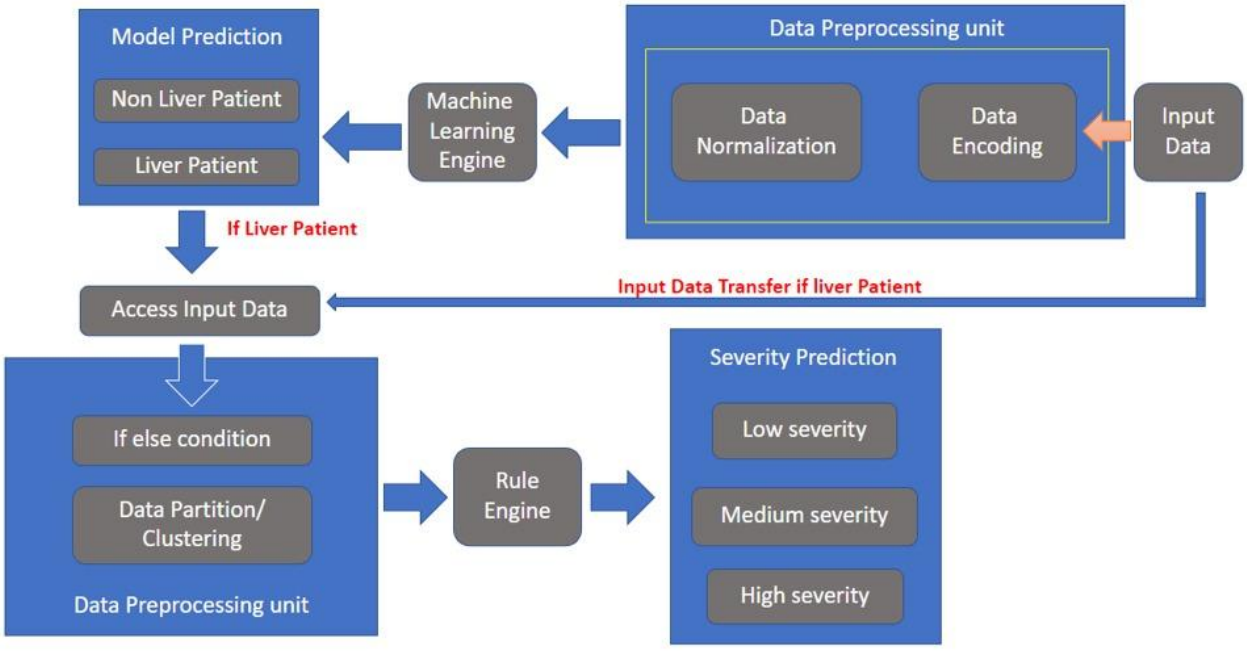

Figure1: System Architecture Diagram

IV. METHODOLOGY

5.1 Research Data

The dataset, known as the Indian Liver Patient Dataset (ILPD)[15], was taken from Kaggle. The target variable is one of 11 features present in 30691 cases in this dataset. Age, gender, direct and total bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, albumin and globulin ratio, and target are the attributes.

5.2 Dataset Analysis and Pre-Processing

The data set includes 8774 patient records who don't have liver disease and 21917 patient records that do.

- Missing Values: Dataset contains 5425 null data points and by dropping null data points we get 27158 instances.
- Data Duplication: Data contains 10769 duplicate rows and by dropping duplicate rows we left with 16389 instances.
- Data Normalization: Data transformation methods like standard scalar are used to normalise data.

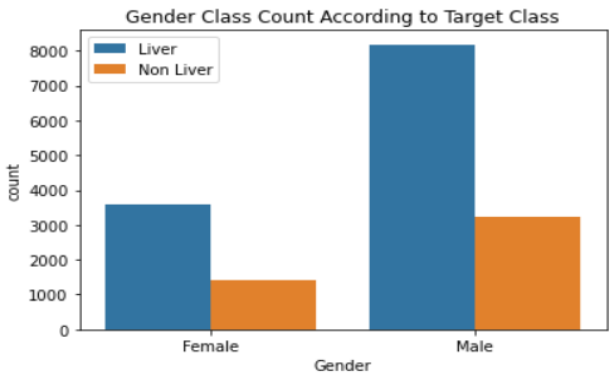

Figure 2: Gender Class count According to Target Class.

We have 11750 liver patients and 4639 non-liver patients after cleansing the data. It includes 4975 records for female patients and 11414 records for male patient. Is shown in Figure: 2

"Target" is a class label with two classes - liver patients (disease) or not (no disease).

5.3 Classifier

1. Decision Tree

A supervised machine learning model called a decision tree can be applied to classification and regression issues. Each interior (non-leaf) hub of a decision tree represents a test on a property, each branch refers to the outcome of a test, and each leaf (or terminal) hub has a class name. The highest hub in a tree is the root hub. There are numerous particular decision tree calculations [17].

2. Support Vector Machine

SVM, or Support Vector Machine, is a tool that can be used for both classification and regression tasks. However, it is frequently employed in classification goals. The support vector machine approach seeks to locate an N-dimensional space hyperplane that clearly categorises the data points. Support Vector Machine is non-probabilistic, so they assign a data point to a class with 100% certainty. [11]

3. Logistic Regression

The probability for classification issues with two possible outcomes are modelled using supervised ML with logistic regression. It is a modification of the linear regression model for issues with classification.[16]. When predicting the binary result for a given collection of independent variables, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1.

4. AdaBoost Classifier

One of the iterative ensemble boosting classifiers is ADA-boost, also known as adaptive boosting. To improve classifier accuracy, it combines several classifiers. AdaBoost classifier combines a number of ineffective classifiers to create a strong classifier that has a high degree of accuracy. The fundamental idea underlying AdaBoost is to train the data sample and set the classifier weights in each iteration to provide accurate predictions of uncommon observations. If a machine learning algorithm accepts weights from the training set, it can be utilised as a base classifier.

5. XGBoost Classifier

Gradient Boosted decision trees are embodied in XGBoost. This technique generates decision trees in a sequential manner. In XGBoost, weights are significant. All independent variables are given weights, which are subsequently used to feed information into the decision tree that forecasts outcomes. Variables that the tree incorrectly predicted are given more weight before being placed into the second decision tree. To create a robust and accurate model, these independent classifiers and predictors are then combined.

V. EXPERIMENTAL RESULT

Six distinct classification techniques, including Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost Classifier (ADB), and XGBoost Classifier, are used to assess the results (XBG). The dataset is split into a training set and a testing set for the experiment. The percentages for the training set are 70% and 30%, respectively. In this study, the machine learning model is trained and tested using 10-fold cross-validation.

A. Performance Parameter

Below is a list of the evaluation measures that are used to gauge how well models function.

1. Confusion matrix

The Confusion Matrix is the tabular representation of actual or predicted values [11]. To assess the parameters, a confusion matrix made up of TP, FP, TN, and FN for actual and predicted data is created.

2. Classification accuracy

The proportion of correctly identified samples to all samples is known as accuracy.

Accuracy = (TP+TN) / (TP+TN+FP+FN)(2)

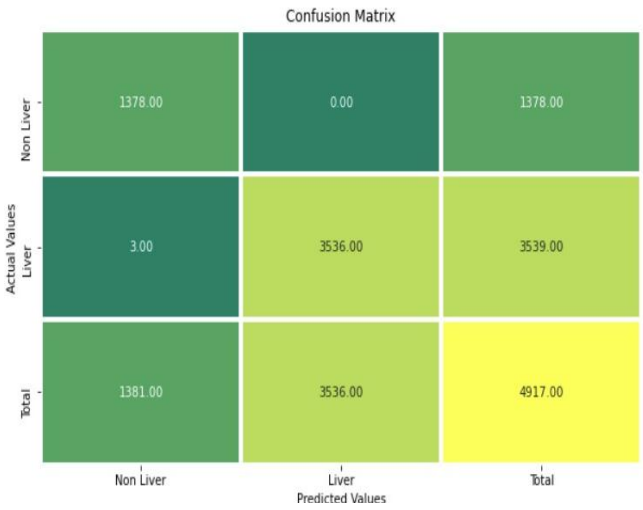

Figure 3: Confusion Matrix of ADB Classifier

3. Recall:

Recall is also viewed as being sensitive. the proportion of actual positive cases to all positive cases. It is true positive rate states how many positive values out of all positive values have been correctly predicted. [11]

Sensitivity=TP/(TP+FN) (3)

4. Precision

Precision is defined as the division of positive cases among all examples that we projected to be positive.

Precision = (TP+TN)/N (4)

5. Area under ROC curve :

The performance of a binary classifier is evaluated using a single scalar metric called the Area Under a Receiver Operating Characteristic (ROC) Curve. The AUC value falls between (0.5-1.0), where the least value corresponds to a random classifier's performance and the maximum value would indicate a perfect classifier. [19]

6. F1 Score :

The **F1-score** combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers. [18]

$F1 = P+R2 / (P*R)$ (5)

7. Kappa Score

Cohen Kappa is an evaluation metric that is often used to measure the degree of agreement between pair of variables, frequently used as a metric of inter-rater (two raters who are rating the same quantity) agreement. Kappa compares the probability of agreement to that expected if the ratings are independent

B. result :

We have train the model on five different classifiers in order to predict whether the person will survive liver disease or not.:

Table 1: Performance measurement parameters for the prediction of Liver disease five using machine learning techniques

	Accuracy	Recall	Precision	F1- score
DT	0.99	1	1	1
LR	0.72	0.55	0.18	0.27
SVM	0.73	0.74	0.05	0.10
ADD	0.78	0.65	0.49	0.56
XGB	0.99	1	1	1

From the above table, we can say that DT and XGB are working properly on the testing dataset with 0.99 of testing Accuracy, But LR, SVM, and ADB have less Accuracy, So hyper parameter tuning is performed on that classifier.

From Figure 4: The accuracy of DT remains the same after performing hyperparameter tuning, but there is an increase in the accuracy of SVC to 0.99 and ADB to 0.85. It contains the selected parameter after performing parameter tuning.

A decision tree is a tree-based algorithm that uses impurity level and information gained for feature importance, therefore many times decision trees overfit according to the features. Whereas AdaBoost is a boosting and sequential machine learning approach that uses weak learners for prediction. AdaBoost by default uses a decision tree with 1 split along. Additionally, it uses a learning rate and weights to focus on non-learned and less signified features of data. First AdaBoost assigns the same weights to all examples and then assigns high weights to the misclassified data to give more importance to the data points in the next iteration. Dual tuning is performed by AdaBoost, on the second iteration accuracy is increased to 99.87%.

The results show that the F1 - Score value is 1, the Precision values are 1, and the Recall value s 0.1

Since in medical terms, test sensitivity is the ability of the test to correctly identify those with the disease thus Ada Boost Classifier is the best model for predicting liver disease. [11]

	Algorithm	Training accuracy	Testing accuracy	AUC Score	Parameter
1	Decision Tree	0.999041	0.999390	0.999576	criterion=entropy,max_depth=25, min_samples_le...
0	SVC	0.997036	0.959935	0.948460	C=1000,kernel-rbf,gamma=1
2	AdaBoost	0.850767	0.838519	0.776154	learning_rate=1.04, n_estimators=5

Figure: 4 Algorithms after performing Parameter tuning along with its selected parameter and Accuracy

After developing a classifier model, AdaBoost Classifier can accurately classify a person's likelihood of having liver illness at 99.87 percent. We have created a rule engine that categorizes liver disease severity into three groups: Low, Normal, and High. Figure 1. The System's design contains the developed loop system's architecture. The process datasets, which are medical parameters used to measure the disease, are entered into the rule engine. Each feature (medical parameter) in the dataset has a range that is determined to divide the severity into three categories.

For group division, a hierarchical model of the if-else conditional ladder is utilized, and an unsupervised clustering algorithm is used to create a cluster of three names: Low, Normal, and High. In order to build a cluster, hierarchical clustering and KMeans were both utilized, with the best outcome coming from hierarchical clustering. The rule engine is a customized rule engine where the outcome can be obtained by manually updating the parametric variables. Data and results from the processing unit are then sent to a custom rule engine, which is managed by the doctor. Depending on patient data and geographic locations, doctors can manually enter attribute values and modify attribute ranges. Therefore, the established rule engine may be evaluated, cross-checked, and validated by the medical professionals. After that, a final yet correct result is generated and put to the test by medical professionals. Finally, severity is forecasted utilizing hybrid machine learning technology. Once a life is saved, which is the most valuable ROI in this paper, it is crucial to predict disease in its early stages and to begin proper treatment.

VI. CONCLUSION

For the purpose of predicting liver illness, various classification methods, including Logistic Regression, Support Vector Machine, Decision Tree Classifier, Ada Boost Classifier, and Xgboost Classifier, have been utilized in this study. These methods have all been compared based on classification accuracy, which is discovered using a confusion matrix. The Decision Tree has the highest accuracy from the experiment, whereas the Ada Boost Classifier has the most sensitivity. AdaBoost Classifier is chosen as the best candidate for predicting liver illness because of its 99.87 percent accuracy and 99.0% sensitivity. Utilizing a rule engine to estimate the severity of a condition based on a set of liver disease symptoms. The data are divided into three groups of three using hierarchical clustering: low, normal, and high. Thus, it can

be said that the severity of liver disease can be anticipated utilizing hybrid machine learning technology.

VII. REFERENCES

[1] Vasan D, Suyan R, Dinesh K "Liver Disease Prediction using machine learning" **** Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] Rakshith D B, Mrigank S, Ashwani K, Gururaj S P "Liver Disease Prediction System using Machine Learning techniques

[3] Md Fazle R, S M Mahedy H, Arifa I C, Md Asif Z, Md Kamrul H " Prediction of Liver Disorder using machine learning algorithm: A comparative study "

[4] Sateesh A, Vijayalaxmi A, Rashmi U, Shruthi P, Vilaskumar P " Optimizing Liver Disease Prediction with random forest by various data balancing techniques"

[5] Javad H, Hamid S, Abdollah D, Shahaboddin S " Computer-aided decision making for predicting liver disease using PSO based optimized SVM with feature selection"

[6] Dr. S. Vijayarani, S. Dhayanand " Liver Disease prediction using SVM and Naïve Bayes Algorithm"

[7] Kalyan N, Amulyashree S " NeuroSVM: a graphical user interface for Identification of Liver Patient"

[8] Miningngpi Z, Changjun S, Tao L, Tianyue H, Shiming L "Fatty Liver Disease Prediction Model Based on Big Data or Electronics Physical Examination records"

[9] Ivair P, Denio D, Guilherme D B, Julyana F L "Exploratory Analysis of Electronic Health Records using Topic Modelling"

[10] Nazim Razali "A data mining approach to the prediction of liver disease"

[11] Thirunavukkarasu K, Ajay S S, Md Irfan, Abhishek C "Prediction of liver disease using classification algorithm"

[12] Fahad M, Easy H, Morgan W, Hafiz K, "Statical Machine Learning Approaches to Liver disease prediction"

[13] Keerthana PSM, Nimish P, Riya M, Koppula Bhanu P R, Nidhi L "A Prediction Model of Detecting Liver disease in a patient using Logistic Regression of Machine Learning"

[14] Insha A, Chiranjit D, Tanupriya C, Abha T "Liver Disease Detection due to excessive alcoholism using data mining techniques"

[15] Available:

https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset.

[16] Available:https://christophm.github.io/interpretable-ml-book/logistic.html

[17] L. Alice Auxilia," Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease".

[18] Available:https://www.educative.io/edpresso/what-is-the-f1-score

[19] Available:https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_209