EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Classification Algorithms for Liver Epidemic Identification

Koteswara Rao Makkena[1], Karthika Natarajan[2,*]

[1,2] School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

Abstract

Situated in the upper right region of the abdomen, beneath the diaphragm and above the stomach, lies the liver is a crucial organ essential for the proper functioning of the body. It performs many important functions in the body. The principal tasks are to eliminate generated waste produced by our organs, and digestive food and preserve vitamins and energy materials. It performs many important functions in the body, It Regulating the balance of hormones in the body Filtering and removing bacteria, viruses, and other harmful substances from the blood. In certain dire circumstances, the outcome can unfortunately result in fatality. There exist numerous classifications of liver diseases, based on their causes or distinguishing characteristics. Some common categories of liver disease include Viral hepatitis, Autoimmune liver disease, Metabolic liver disease, Alcohol-related liver disease, Non-alcoholic fatty liver disease, Genetic liver disease, Drug-induced liver injury, Biliary tract disorders. Machine learning algorithms can help identify patterns and risk factors that may be difficult for humans to detect. With This clinician can enable early diagnosis of diseases, leading to better treatment outcomes and improved patient care. In this research work, different types of machine learning methods are implemented and compared in terms of performance metrics to identify whether a person effected or not. The algorithms used here for predicting liver patients are Random Forest classifier, K-nearest neighbor, XGBoost, Decision tree, Logistic Regression, support vector machine, Extra Trees Classifier. The experimental results showed that the accuracy of various machine learning models-Random Forest classifier-67.4%, K-nearest neighbor-54.8%, XGBoost-72%, Decision tree-65.1%, Logistic Regression-68.0%, support vector machine-65.1%, Extra Trees Classifier-70.2% after applying Synthetic Minority Over-sampling technique.

Keywords: medical care, liver epidemic, prognosis, classification models, Synthetic Minority Over -sampling technique

Received on 10 September 2023, accepted on 07 November 2023, published on 13 November 2023

doi: 10.4108/eetpht.9.4379

*Corresponding author. Email: karthika.n@vitap.ac.in

1. Introduction

Nestled on the right side of the abdomen, slightly below the diaphragm, resides the liver—a sizable, reddish-brown organ. It performs a variety of vital functions in the body like metabolism, detoxification, synthesis, storage, and immunity. This is why some liver conditions can even be fatal [1,2].

There are different types of liver epidemic based on their underlying causes, such as viral infections (hepatitis A, B, C, etc.) [3], alcohol abuse [4], nonalcoholic fatty liver disease [5], genetic disorders, autoimmune diseases, and drug-induced liver injury. Liver diseases can be categorized based on the effects they have on the liver's function, such as cirrhosis, hepatitis, or liver cancer [6]. Cirrhosis is a progressive liver disease resulting from long-term liver damage, such as alcohol abuse or chronic viral hepatitis. It causes the liver tissue to become scarred, leading to a loss of liver function [7].

Hepatitis refers to the inflammation of liver tissue. Prolonged inflammation, known as chronic hepatitis, can result in the formation of scar tissue (fibrosis), which may progress to irreversible scarring (cirrhosis) and even hepatocellular carcinoma (HCC). The liver can undergo inflammation due to various factors, including alcohol

consumption, certain medications, autoimmune disease, fatty liver disease, and viral infections. [3].

Hepatitis A is an infectious liver infection triggered by the hepatitis A virus (HAV). The HAV can be found in the blood and stool of infected individuals. The transmission of the hepatitis A virus occurs primarily through person-to-person contact and the consumption of contaminated food or beverages. [8].

Hepatitis B is a viral infection that affects the liver. It is caused by the hepatitis B virus (HBV), which is transmitted through contact with infected blood, semen, or other bodily fluids. Hepatitis B can lead to both acute and chronic liver disease, and can cause serious health problems if left untreated [9].

The pattern for hepatitis C is the same as for the first two. However, it passes on through getting contacting an infected person's blood. Even ten years after infection, early symptoms can appear. As there are the two acute and chronic different types of hepatitis B [10].

Only patients with hepatitis B are subject to the very uncommon hepatitis D virus [11], and by drinking water that has been contaminated by an infected person's waste, hepatitis E is able to spread [12].

While hepatitis or alcoholism are usually causes of liver disease, a major threat of likely fatal liver damage is caused by obesity and diabetes. During the advanced stages of fatty liver disease, the liver can undergo significant damage, impairing its proper functioning. This can result in a variety of grave complications, such as liver failure, liver cancer, and an elevated risk of mortality [13-15].

Patients should not consume a lot of alcohol to avoid liver disease. However, the straightforward advice for individuals who have been diagnosed with hepatitis B or C, alcoholic hepatitis, etc. is to never drink alcohol at all. Use of a condom during sexual activity, avoid from sharing needles or syringes, being vaccinated against hepatitis A and B, and protecting the skin from toxins are further measures [16-17].

Traditionally, healthcare professionals base their interpretations of a patient's condition on histological studies. Effective approaches for data gathering, analysis, and visualization have emerged because of the advanced Technologies like Intelligent automation and automated learning.

Clinicians might further enhance their decisions on disease detection by integrating the findings of Intelligent automation and automated learning models with those of clinical approaches [62]. The quick identification of disease complications in diabetes has greatly benefited from the use of machine learning techniques [18-19] or regression tasks for "short-term blood glucose forecasting [20], lipid [21], high blood pressure [22], high cholesterol [23], chronic inflammatory lung disease (COPD) [24], novel coronavirus disease [25], cerebrovascular accident [26], chronic renal disease [27], pulmonary carcinoma [28], insomnia [29], coronary artery disease [30].

In the context of this scientific research, the occurrence of liver disease will be of special significance to us. The following are the main contributions of the chosen methodology:

- The Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling approach that creates synthetic minority class samples [61]. With SMOTE technique imbalanced dataset is balanced allowing to design of machine learning models and identify whether a person effected with liver disease or not [61].
- Correlation among attributes can be specified based on the Pearson's coefficient and XGBoost classifier.
- Prominent metrics like accuracy, precision, recall, F1 score, and AUC are commonly employed to assess the performance of numerous machine learning models.

2. Motivation and objectives

2.1 Motivation:

The motivation behind using supervised machine learning models for liver disease risk prediction is to enhance early detection and provide timely medical interventions [62].

Liver diseases, including conditions like cirrhosis [63] and hepatitis [64], can be life-threatening if not diagnosed and treated in their early stages.

By leveraging machine learning algorithms, we can analyze patient data and identify patterns and risk factors associated with liver diseases.

2.2 Objectives:

The objectives of using supervised machine learning models for liver disease risk prediction include:

Early Detection: The primary objective is to develop accurate models that can identify individuals at risk of developing liver diseases at an early stage. By leveraging patient data, including medical history, laboratory test results, and demographic information, these models can identify patterns and risk factors associated with liver diseases.

Risk Assessment: Another objective is to assess the risk level of individuals for developing liver diseases. Machine

learning models can assign a risk score or probability that indicates the likelihood of an individual developing a liver disease within a certain timeframe. This information can help prioritize patients for further diagnostic tests or targeted interventions.

Feature Importance: Understanding the importance of various features or risk factors associated with liver diseases is crucial for medical professionals. By leveraging machine learning models, it becomes possible to gain valuable insights into the variables that wield substantial influence on risk prediction. This knowledge can guide medical practitioners in focusing on the most relevant factors during diagnosis and treatment planning.

Model Comparison and Selection: Comparing different supervised machine learning models, such as logistic regression, decision trees, random forest, support vector machines (SVM), and gradient boosting algorithms like XGBoost, is an essential objective [62,65]. By evaluating and comparing the performance of these models, we can identify the most effective approach for liver disease risk prediction.

Model Interpretability: In addition to accurate predictions, it is desirable to have models that are interpretable. Interpretable models provide insights into the decision-making process, allowing medical professionals to understand the underlying reasons behind the risk predictions. This can help build trust and acceptance of the model among healthcare practitioners.

3. Background and Techniques

Two datasets are used and the key components of the methodology we used to forecast the risk of liver disease, namely imbalanced data, and Attributes selection in the balanced data.

3.1 Data Set-1 Description

Research is based on dataset 1 [31]. In dataset 1 Gender feature can be represent in terms numbers 3(male) ,4(female). This dataset contains 583 people, of which the number of 3(441) and 4 (142) also unbalanced dataset. Data set 1 has some missing values on specific Attribute. These missing indent.

This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent.

Table 1: dataset-1 attributes

Sno	Attribute	Data type	Explanation
1	Age [32]	number	Age range is in b/w 4–90 years.
2	Gender [33]	character	Attribute illustrates the people's gender.
3	TBIL (mg/dL) [34]	number	Attribute specifies the people's TBIL.
4	DBIL [34]	number	Attribute specifies the people's DBIL.
5	AP [35]	number	Attribute illustrates the people's AP
6	Alanine transaminase [36]	number	Attribute specifies the people's Alanine transaminase
7	Aspartate transaminase [36]	number	Attribute specifies the people's Aspartate transaminase
8	Total_Protiens -TP(g/L) [37]	number	Attribute specifies the people's total protein.
9	Aluminosol (g/L) [38]	number	Attribute specifies the people's Aluminosol
10	AGR test [39]	number	Attribute specifies the people's AGR
11	outcome	number	outcome.

3.2 Data Set-2 Description

Research is based on dataset 2[40]. In dataset 2 Gender feature can be represent in terms numbers 1(male) ,0(female) It contains 117 peoples, of which the number of 1 is 98 (68.4%) and 0 is 18 (31.6%) also balanced datasets. This dataset contains no missing values. The target class is the outcome which indicates whether the people have a liver disease or not. Dataset characteristics can be viewed below in Table 2.

Table 2: Dataset-2 Description

Sno	Attribute	Type	Description
1	Age [32]	number	age range is 4–90 years.
2	Gender [33]	character	Attribute illustrates the people gender.
3	TBIL (mg/dL) [34]	number	Attribute specifies the people's TBIL
4	DBIL [34]	number	Attribute specifies the people's DBIL.
5	IDBIL [34]	number	Attribute specifies the people's IDBIL
6	Alanine transaminase [35]	number	Attribute illustrates the people's Alanine transaminase.
7	Glutamate-pyruvate transaminase [36]	number	Attribute specifies the people's Glutamate-pyruvate

			transaminase
8	Glutamic-oxaloacetic transaminase [36]	number	Attribute specifies the people's Glutamic-oxaloacetic transaminase.
9	Total_Protiens-TP(g/L) [37]	number	Attribute specifies the people's total protein.
10	Globulin (g/L) [38]	number	Attribute specifies the people's Globulin.
11	Aluminosol (g/L) [38]	number	Attribute specifies the people's Aluminosol
12	AGR test [39]	number	Attribute specifies the people's albumin and g lobulin Ratio.
13	class	number	class

3.3 Liver epidemic risk calculator

Today medical experts use machine learning models to create effective tools to evaluate the possibility of a disease occurring based on different causes. If a person is affected or not is represented as a ML model with two possible values of attribute outcome either binary values 1 or 0. If Attribute outcome value is 1 then patient has a liver disease otherwise not diseased.

3.3.1. Data Pre-processing

The unequal distribution of outcome Attribute in the dataset may affect how accurately such instances are identified. An oversampling method SMOTE [41] applied for balancing the data set based on the KNN classifier creates simulated data [42]. Following the application of SMOTE on Dataset 1, the dataset now comprises a total of 400 individuals, with 288 (72%) identified as men (male) and 120 (30%) identified as women (female), resulting in a balanced distribution. Consequently, the target class consists of 288 instances labelled as 1 and 288 instances labelled as 0. Finally, the following Table 2, shows that statistical summary of the various attributes in balanced dataset.

Table 3: Statistical summary of different attributes

Attribute	Min	max	mean±stdv
age	4	90	44.74±16.18
tb	0.4	75	3.298±6.2
DB	0.1	19.7	1.48±2.808
ALP	63	2110	290.57±242.93
SGPT	10	2000	80.71±182.62
SGOT	10	4929	109.91±288.91
tp	2.7	9.6	6.48±1.08
alb	0.9	5.5	3.14±0.79
agr	0.3	2.8	0.94±0.31

3.3.2 feature analysis:

One of the ranking methods [43] has been selected to evaluate the contribution of an Attribute in the outcome (1) in data set 1 & class in dataset 2. Their results are illustrated in Table 3.

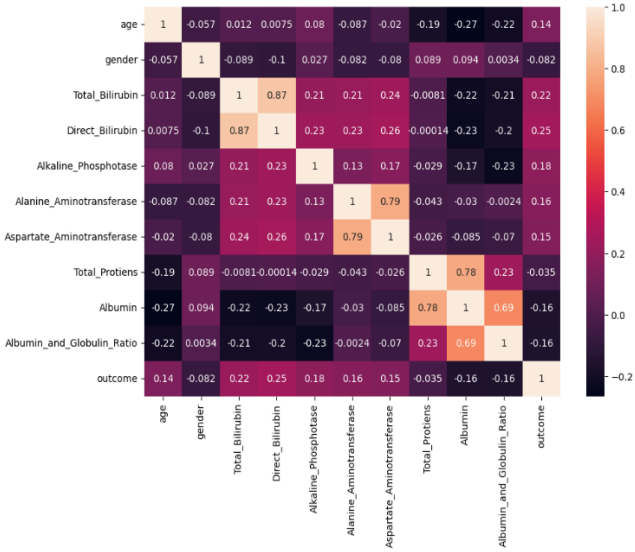


Figure 1: Pearson correlation analysis can be specified by heatmap.

Table 4. Attribute importance is classified based on the XGBoost Classifier for Dataset 1[31]

Attribute	Value
Direct Bilirubin	0.181159
Total Bilirubin	0.172055
Albumin	0.093506
Alanine Aminotransferase	0.090644
age	0.086543
Alkaline_Phosphotase	0.084206
Total_Protiens	0.084206
Albumin_and_Globulin_Ratio	0.073368
gender	0.072695
Aspartate_Aminotransferase	0.067070

While implementing XGBoost Classifier for Attribute importance, Direct Bilirubin and Total Bilirubin can show in the below figure 2.

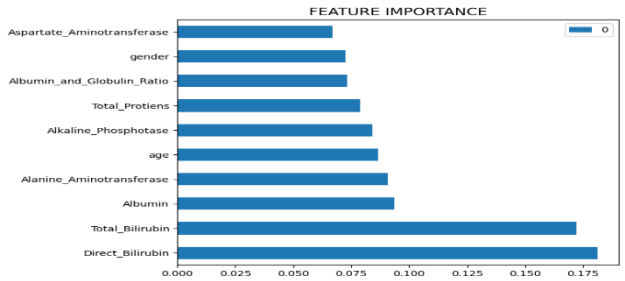


Figure 2. Shows Attribute importance for the analysis of Attributes in our data set 1

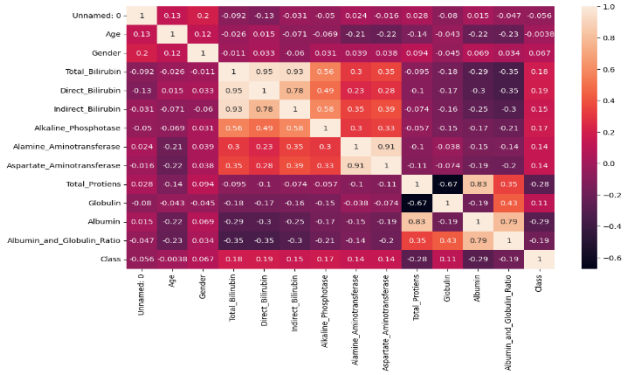


Figure 3: Pearson correlation analysis can be specified by heatmap

Table 5. Attribute importance is classified based on XGBoost Classifier for Dataset 2[40]

Attribute	value
Indirect_Bilirubin	0.166936
Aspartate_Aminotransferase	0.150382
Direct_Bilirubin	0.124006
Total_Bilirubin	0.086991
Alkaline_Phosphotase	0.080814
Albumin	0.080168
Albumin_and_Globulin_Ratio	0.055506
Globulin	0.054525
Total_Protiens	0.051350
Age	0.047039
Alamine_Aminotransferase	0.046970
Gender	0.013968

While implementing XGBoost Classifier for Attribute importance, Indirect_Bilirubin and Aspartate_Aminotransferase can show in the below figure 3.

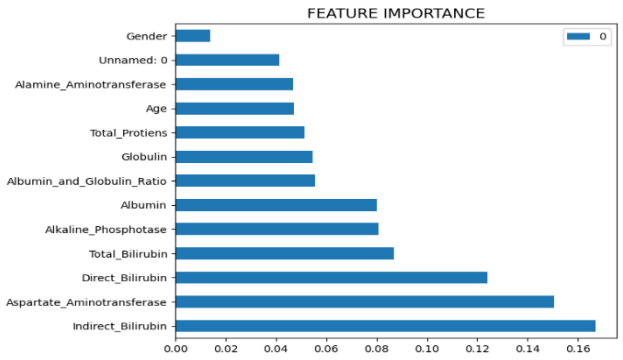


Figure 4. Shows Attribute importance for the analysis of Attributes in our data set 1

3.4 Discriminative models:

During this discussion, several types of machine learning (ML) models were implemented to determine their relative performance based on accuracy. A particular emphasis was placed on Logistic Regression [44], known for its probabilistic classification approach. Furthermore, a kernel-based Support Vector Machine (SVM) was also employed in the analysis [45]. We used the Decision tree algorithm [46], k-nearest neighbours (KNN) [47], and Ensemble learning algorithms like Extra Tree classifier [48], Random Forest algorithm [49], and XGBoost classifier [50] are evaluated.

3.5. Evaluation Metrics:

Machine learning models' performance is evaluated by using common metrics like Accuracy, Precision, Recall, F-Measure, and AUC [51-52].

- Accuracy: Accuracy refers to the measure of correctly classified data instances in relation to the total number of data instances.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

- Recall: Recall measures the model ability to detect positive samples.

$$Recall = \frac{TP}{TP + FN}$$

- Precision: Precision is a statistical metric employed to assess the precision or accuracy of predictions made by a model. It is the percentage

of true positives (i.e., correct predictions) out of all positive predictions (i.e., all predictions that the model made for that class).

$$Precision = \frac{TP}{TP + FP}$$

- F-Measure: The F1 Score attains a value of 1 only when both precision and recall reach 1. A high F1 Score is achieved when both precision and recall are high. Serving as the harmonic mean of precision and recall, the F1 Score is considered a superior measure compared to accuracy.

$$F - Measure = 2\frac{Precision.Recall}{Precision + Recall}$$

The AUC is used to assess a model's ability to be identified. It is a variable measure between [0, 1].

4. Findings:

4.1 Performance assessment based on accuracy:

The performance evaluation of different classification models often relies on accuracy as a metric. Presented below are two tables showcasing the performance of seven ML models before and after applying SMOTE to dataset 1 and dataset 2.

Table 6: Accuracy score of 7 classification models dataset 1

Sno	classification models	accuracy score
1	KNN	0.668571
2	XGBoost	0.760000
3	Random Forest	0.748571
4	Decision Tree	0.674286
5	**Logistic Regression**	**0.771429**
6	ExtraTress Classifier	0.737143
7	Support Vector Machine	0.737143

Before applying SMOTE technique, logistic regression classification model has high accuracy score.

Table 7: Efficiency of ML models after applying SMOTE technique on dataset 1

sno	classification models	accuracy score
1	KNN	0.548571
2	**XGBoost**	**0.720000**
3	Random Forest	0.674286
4	Decision Tree	0.651429
5	Logistic Regression	0.680000
6	ExtraTrees Classifier	0.702857
7	Support Vector Machine	0.651429

After applying SMOTE technique on dataset 1, XGBoost classification model has high accuracy score.

Table 8: Performance of 7 ML models before SMOTE on dataset 2

Sno	Classification Models	Accuracy_score
1	KNN	0.805556
2	XGBoost	0.805556
3	Random Forest	0.833333
4	Decision Tree	0.777778
5	Logistic Regression	0.861111
6	**ExtraTrees**	**0.888889**
7	Support vector Machine	0.861111

Before applying SMOTE technique on dataset 2

ExtraTrees classifier has highest accuracy score.

Table 9: Performance of ML models after applying SMOTE technique on dataset 2.

Sno	classification models	Accuracy_score
1	KNN	0.722222
2	**XGBoost**	**0.861111**
3	**Random Forest**	**0.861111**
4	Decision Tree	0.777778
5	Logistic Regression	0.722222
6	ExtraTrees	0.805556
7	Support vector Machine	0.722222

After apply SMOTE technique on dataset 2 both XGBoost Classifier and Random Forest classifier has highest accuracy score.

XGBoost classifier and Random Forest classifiers often yield better results compared to other models for several reasons:
Both XGBoost and Random Forest are ensemble methods that combine multiple individual models (decision trees in this case) to make predictions. Ensemble methods have the advantage of reducing bias, variance, and overfitting by aggregating the predictions of multiple models.
XGBoost and Random Forest can capture nonlinear relationships between features and the target variable. They can model complex interactions and capture nonlinearity more effectively than linear models.

Both models are robust to outliers in the dataset. Outliers can significantly affect the performance of linear models, but decision tree-based models like XGBoost and Random Forest can handle them better.

These models provide measures of feature importance, allowing you to identify the most influential features in the prediction process.

Both XGBoost and Random Forest have several hyperparameters that can be fine-tuned to optimize their performance.
Tuning these parameters through techniques like grid search or random search can further improve the models' accuracy.

The following figures represent the accuracy score for 7 classification models before and after applying SMOTE technique on dataset 1 and dataset 2.

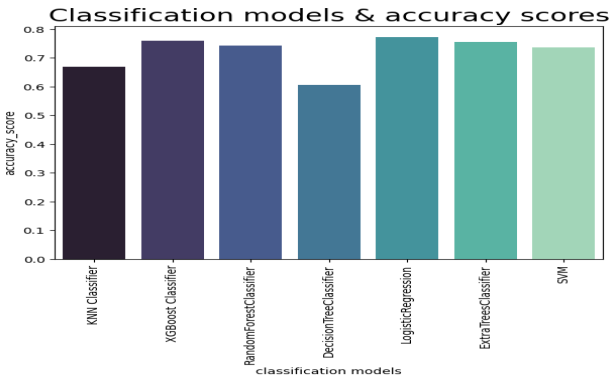


Figure 5: Accuracy scores for 7 classification models on dataset 1 before apply SMOTE

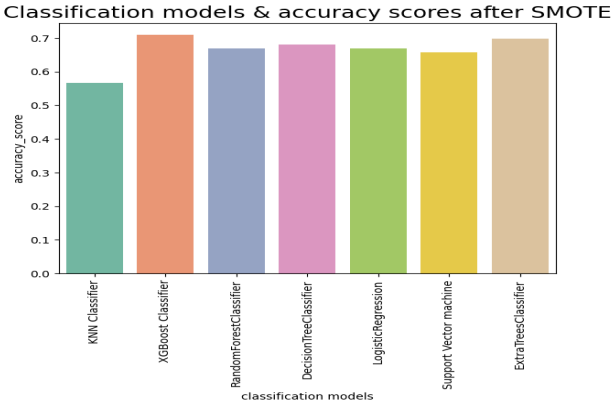


Figure 6: Accuracy scores for 7 classification models on dataset 1 after apply SMOTE

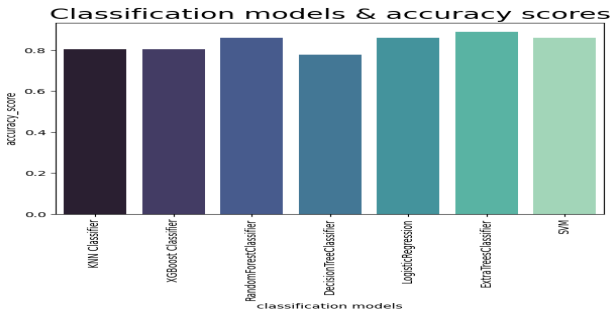


Figure 7: Accuracy scores for 7 classification models on dataset 2 before apply SMOTE

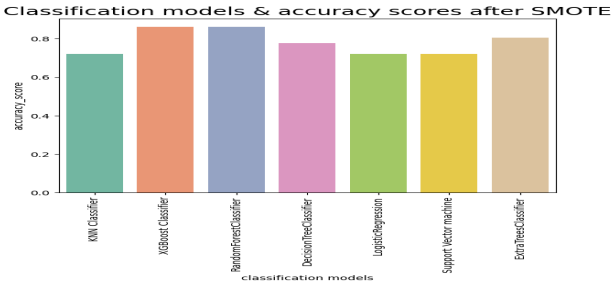
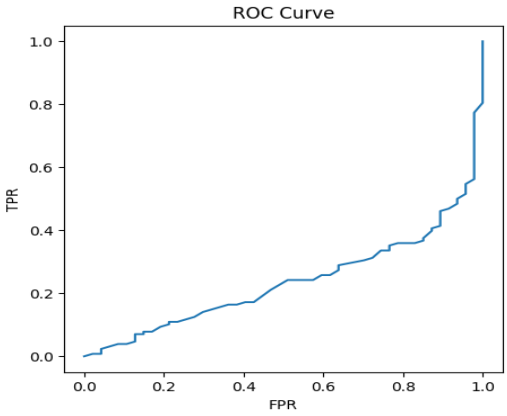


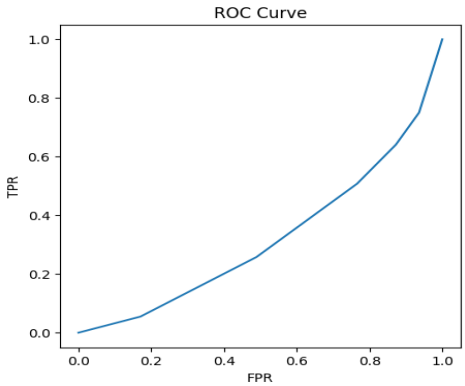
Figure 8: Accuracy scores for 7 classification models on dataset 2 after apply SMOTE

Along with evaluation metrics, we proved the model improvement by AUC ROC Curve after applying SMOTE technique.AUC-ROC curve is used to visualize the performance of a classification model. The AUC (Area Under the Curve) value varies between 0 and 1, with a higher value indicating a better model. An excellent model

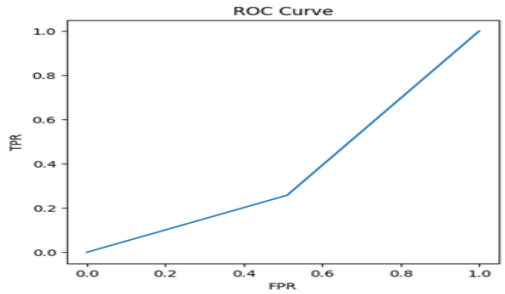
will exhibit an AUC close to 1, signifying a strong measure of separability.



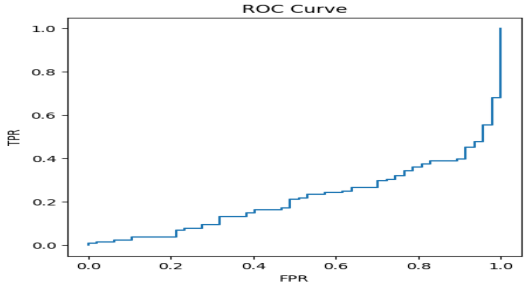
RT



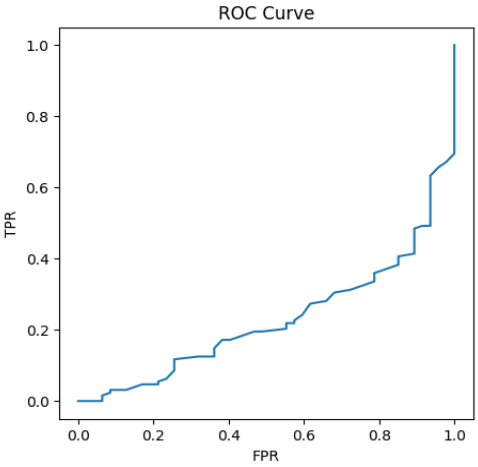
KNN



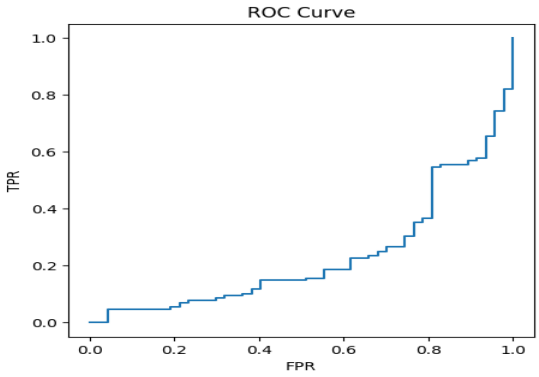
Decision Tree



Logistic Regression



ExtraTrees classifier



XGBoost classifier

Figure 9: ROC-AUC Curves for classification models after apply SMOTE technique on data set1

Considering Decision Tree classifier for ROC &AUC as this model showing better accuracy among remaining classification models used on data set1.The following figure shows various classification models after applying SMOTE technique on data set 2.

4.2 Evaluation of The Model:

Performance of the various machine learning models can also measure in terms of precision, recall, f1-score, and support. The following tables represents various machine learning models performance.

4.2.1 Random Forest classifier:

Table 10: Performance of RF model after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.43	0.91	0.59	47
1	0.95	0.56	0.71	128
Accuracy			0.66	175
Macro Avg	0.69	0.74	0.65	175
Weighted Avg	0.81	0.66	0.67	175

4.2.2 Random Forest classifier:

Table 11: Performance of RF model after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.43	0.91	0.59	47
1	0.95	0.56	0.71	128
Accuracy			0.66	175
Macro Avg	0.69	0.74	0.65	175
Weighted Avg	0.81	0.66	0.67	175

4.2.3 KNN Classifier:

Table 12: Performance of KNN model after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.36	0.77	0.49	47
1	0.85	0.49	0.62	128
Accuracy			0.57	175
Macro Avg	0.60	0.63	0.56	175
Weighted Avg	0.72	0.57	0.59	175

4.2.4 XGBoost Classifier:

Table 13: Performance of XGBoost model after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.49	0.55	0.52	47
1	0.83	0.79	0.81	128
Accuracy			0.73	175
Macro Avg	0.66	0.67	0.66	175
Weighted Avg	0.74	0.73	0.73	175

4.2.5 Decision Tree Classifier:

Table 14: Performance of Decision Tree model after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.37	0.40	0.38	47
1	0.77	0.74	0.76	128
Accuracy			0.65	175
Macro Avg	0.57	0.57	0.57	175
Weighted Avg	0.66	0.65	0.66	175

4.2.6 Logistic Regression:

Table 15: Performance of Logistic regression after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.46	0.89	0.59	47
1	0.94	0.61	0.71	128
Accuracy			0.69	175
Macro Avg	0.70	0.75	0.67	175
Weighted Avg	0.81	0.69	0.70	175

4.2.7 Support vector machine:

Table 16: Performance of Support vector machine models after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.43	0.91	0.59	47
1	0.95	0.56	0.71	128
Accuracy			0.66	175
Macro Avg	0.69	0.74	0.65	175
Weighted Avg	0.81	0.66	0.67	175

4.2.8 Extraa Trees Classifier:

Table 17: Performance of Extraa Trees model after applying SMOTE technique on dataset 1.

	Precision	Recall	F1-score	Support
0	0.51	0.64	0.57	47
1	0.85	0.77	0.81	128
Accuracy			0.74	175
Macro Avg	0.68	0.71	0.69	175
Weighted Avg	0.76	0.74	0.75	175

5. Discussion:

Liver epidemic based on two data sets dataset 1 [31] and dataset 2 [40] are presented to check whether a person is affected or not by applying various machine learning models [53]. Specifically, the KNN classifier model [55] reached an accuracy (of 66.85%), the XGBoost classifier model reaching accuracy (76%), Random Forest classifier model [54] reaching accuracy (74.85%).

In addition, Decision Tree classifier model [56] reaching accuracy (67.42%), logistic Regression classifier model [57] reaching accuracy (77.14%), ExtraTrees Classifier model [58] reaching accuracy (73.71%) and support vector machine classifier model [59] reaching an accuracy of (73.71%) before applying SMOTE technique on Indian Liver Patient Dataset.

The various classification models get accuracy after implementing SMOTE technique for balancing the dataset. These are the KNN classifier model [55] reached an accuracy (of 54.85%), the XGBoost classifier model reaching an accuracy (72%), Random Forest classifier model reaching an accuracy (67.42%). In addition, Decision Tree classifier model reaching accuracy (65.14%), the logistic Regression classifier model reaching accuracy (68%), ExtraTrees Classifier model reaching accuracy (70.28%) and support vector machine classifier model reaching accuracy of (65.14%).

Attribute importance can be chosen by applying XGBoost ensemble learning method to selecting which Attributes are important on both data sets. For Indian Liver Patient Dataset Total Bilirubin, and Direct Bilirubin are the two important Attributes while prediction of liver disease. While applying XGBoost ensemble learning model on Indian Liver Patient Records Dataset Indirect_Bilirubin, Aspartate_Aminotransferase are two important Attributes for prediction of liver disease.

We focussed on performance evaluation based on graphical specifies of AUC -ROC curves on two datasets. While applying AUC-ROC on Indian Liver Patient Dataset [31], decision tree classifier model got highest accuracy among the remaining classification models. The same AUC-ROC apply on another dataset Indian Liver Patient Records [40], decision tree classifier model got high accuracy Performance is also calculate based on precision, recall, f1-score, support for various machine learning models.

The results of this study will contribute to liver disease monitoring and facilitate the creation of customized models with exceptional performance. These models will possess the flexibility to incorporate both quality-of-life aspects that illuminate the challenges associated with this condition and reflect the well-being of patients.

Finally, using data from a medical center or institute might present us with a wider range of Attributes to properly assess the ML models.

6. Conclusion & Future Scope

The presence of a severe liver illness raises a critical concern that requires immediate medical intervention. Healthcare providers utilize pathological procedures to generate medical reports on a patient's condition. This study aimed to predict early liver disease utilizing machine learning methods.

In particular, a variety of machine learning models, such as KNN, XGBoost, Random Forest, Decision Tree, Logistic Regression, SVM, and ExtraTrees, were examined using evaluation metrics to forecast the onset of liver disease.

The results indicate the accuracy of various classification models both before and after applying SMOTE techniques to two datasets. Additionally, the AUC-ROC curve was employed to visually represent the performance evaluation of the classification models after applying the SMOTE

technique along with precision, recall, f1-score, support are measures for calculation of performance of various ML models.

In future research, the scope will be expanded to include additional machine learning models, incorporating deep learning methods, and comparing the outcomes based on performance metrics.

Acknowledgements.

I extend my heartfelt gratitude to all who played a vital role in the completion of our liver epidemic identification classification algorithms paper. Special thanks to healthcare professionals for sharing valuable data now publicly available on Kaggle. I am especially thankful to my research supervisor, Dr. Karthika Natarajan, for her expertise, guidance, and unwavering support, shaping this study.

References

[1] Arias, I.M.; Alter, H.J.; Boyer, J.L.; Cohen, D.E.; Shafritz, D.A.; Thorgeirsson, S.S.; Wolkoff, A.W. The Liver: Biology and Pathobiology;John Wiley & Sons: Hoboken, NJ, USA, 2020.

[2] Singh, H.R.; Rabi, S. Study of morphological variations of liver in human. Transl. Res. Anat. 2019, 14, 1–5.

[3] Razavi, H. Global epidemiology of viral hepatitis. Gastroenterol. Clin. 2020, 49, 179–189.

[4] Seitz, H.K., Bataller, R., Cortez-Pinto, H. et al. Alcoholic liver disease. Nat Rev Dis Primers 4, 16 (2018). https://doi.org/10.1038/s41572-018-0014-7

[5] Powell, E.E.;Wong, V.W.S.; Rinella, M. Non-alcoholic fatty liver disease. Lancet 2021, 397, 2212–2224.

[6] Ringehan, M.; McKeating, J.A.; Protzer, U. Viral hepatitis and liver cancer. Philos. Trans. R. Soc. B Biol. Sci. 2017, 372, 20160274.

[7] Smith, A.; Baumgartner, K.; Bositis, C. Cirrhosis: Diagnosis and management. Am. Fam. Physician 2019, 100, 759–770.

[8]https://www.cdc.gov/hepatitis/hav/pdfs/HepAGeneralFactSheet.pdf

[9] Yuen, M.F.; Chen, D.S.; Dusheiko, G.M.; Janssen, H.L.; Lau, D.T.; Locarnini, S.A.; Peters, M.G.; Lai, C.L. Hepatitis B virus infection.Nat. Rev. Dis. Prim. 2018, 4, 1–20.

[10] Manns, M.P.; Buti, M.; Gane, E.; Pawlotsky, J.M.; Razavi, H.; Terrault, N.; Younossi, Z. Hepatitis C virus infection. Nat. Rev. Dis.Prim. 2017, 3, 1–19.

[11] Mentha, N.; Clément, S.; Negro, F.; Alfaiate, D. A review on hepatitis D: From virology to new therapies. J. Adv. Res. 2019,17, 3–15.

[12] Kamar, N.; Izopet, J.; Pavio, N.; Aggarwal, R.; Labrique, A.;Wedemeyer, H.; Dalton, H.R. Hepatitis E virus infection. Nat. Rev.Dis. Prim. 2017, 3, 1–16.

[13] Marchesini, G.; Moscatiello, S.; Di Domizio, S.; Forlani, G. Obesity-associated liver disease. J. Clin. Endocrinol. Metab. 2008,93, s74–s80.

[14] Seitz, H.K.; Bataller, R.; Cortez-Pinto, H.; Gao, B.; Gual, A.; Lackner, C.; Mathurin, P.; Mueller, S.; Szabo, G.; Tsukamoto, H.Alcoholic liver disease. Nat. Rev. Dis. Prim. 2018, 4, 1–22.

[15] Åberg, F.; Färkkilä, M. Drinking and obesity: Alcoholic liver disease/nonalcoholic fatty liver disease interactions. In Seminars in Liver Disease; Thieme Medical Publishers: New York, NY, USA, 2020; Volume 40, pp. 154–162.

[16] Bae, M.; Park, Y.K.; Lee, J.Y. Food components with antifibrotic activity and implications in prevention of liver disease. J. Nutr.Biochem. 2018, 55, 1–11.

[17] Cai, J.; Zhang, X.J.; Li, H. Progress and challenges in the prevention and control of nonalcoholic fatty liver disease. Med. Res. Rev.2019, 39, 328–348.

[18] Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access 2021, 9, 103737–103757.

[19]. Dritsas, E.; Trigka, M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors 2022, 22, 5304.

[20] Alexiou, S.; Dritsas, E.; Kocsis, O.; Moustakas, K.; Fakotakis, N. An approach for Personalized Continuous Glucose Prediction with Regression Trees. In Proceedings of the 2021 6th South-East Europe Design Automation, Computer Engineering, ComputerNetworks and Social Media Conference (SEEDA-CECNSM), Preveza, Greece, 24–26 September 2021; pp. 1–6.

[21].Fazakis, N.; Dritsas, E.; Kocsis, O.; Fakotakis, N.; Moustakas, K. Long-Term Cholesterol Risk Prediction with Machine Learning Techniques in ELSA Database. In Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI), Online, 24–26 October 2021; pp. 445–450.

[22] Dritsas, E.; Fazakis, N.; Kocsis, O.; Fakotakis, N.; Moustakas, K. Long-Term Hypertension Risk Prediction with ML Techniques in ELSA Database. In Proceedings of the International Conference on Learning and Intelligent Optimization, Athens, Greece, 20–25 June 2021; pp. 113–120.

[23] Dritsas, E.; Trigka, M. Machine Learning Methods for Hypercholesterolemia Long-Term Risk Prediction. Sensors 2022, 22, 5365.

[24] Dritsas, E.; Alexiou, S.; Moustakas, K. COPD Severity Prediction in Elderly with ML Techniques. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu Island, Greece, 29 June–1 July 2022; pp. 185–189.

[25] Dritsas, E.; Trigka, M. Supervised Machine Learning Models to Identify Early-Stage Symptoms of SARS-CoV-2. Sensors 2023,3, 40.

[26] Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques.

[27] Dritsas, E.; Trigka, M. Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. Big Data Cogn. Comput. 2022, 6, 98.

[28] Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn. Comput. 2022, 6, 1

[29] Konstantoulas, I.; Kocsis, O.; Dritsas, E.; Fakotakis, N.; Moustakas, K. Sleep Quality Monitoring with Human Assisted Corrections. In Proceedings of the International Joint Conference on Computational Intelligence (IJCCI), Online, 24–26 October 2021; pp. 435–444.

[30] Dritsas, E.; Alexiou, S.; Moustakas, K. Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. In Proceedings of the ICT4AWE, Online, 23–25 April 2022; pp. 315–321.

[31. Indian Liver Patient Records. Available online: https://www.kaggle.com/datasets/uciml/indian-liver-patient-records (accessed on 14 November 2022).

[32]Lin, H.; Yip, T.C.F.; Zhang, X.; Li, G.; Tse, Y.K.; Hui, V.W.K.; Liang, L.Y.; Lai, J.C.T.; Chan, S.L.; Chan, H.L.Y.; et al. Age and the relative importance of liver-related deaths in nonalcoholic fatty liver disease. Hepatology 2022.

[33] Mauvais-Jarvis, F.; Merz, N.B.; Barnes, P.J.; Brinton, R.D.; Carrero, J.J.; DeMeo, D.L.; De Vries, G.J.; Epperson, C.N.; Govindan, R.; Klein, S.L.; et al. Sex and gender: Modifiers of health, disease, and medicine. Lancet 2020, 396, 565–582.

[34] Ruiz, A.R.G.; Crespo, J.; Martínez, R.M.L.; Iruzubieta, P.; Mercadal, G.C.; Garcés, M.L.; Lavin, B.; Ruiz, M.M. Measurement, and clinical usefulness of bilirubin in liver disease. Adv. Lab. Med. Med. Lab. 2021, 2, 352–361.

[35] Liu, Y.; Cavallaro, P.M.; Kim, B.M.; Liu, T.; Wang, H.; Kühn, F.; Adiliaghdam, F.; Liu, E.; Vasan, R.; Samarbafzadeh, E.; et al. A role for intestinal alkaline phosphatase in preventing liver fibrosis. Theranostics 2021, 11, 14.

[36] Goodarzi, R.; Sabzian, K.; Shishehbor, F.; Mansoori, A. Does turmeric/curcumin supplementation improve serum alanine aminotransferase and aspartate aminotransferase levels in patients with nonalcoholic fatty liver disease? A systematic review and meta-analysis of randomized controlled trials. Phytother. Res. 2019, 33, 561–570.

[37] He, B.; Shi, J.;Wang, X.; Jiang, H.; Zhu, H.J. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. BMC Biol. 2020, 18, 1–16.

[38] Carvalho, J.R.; Machado, M.V. New insights about albumin and liver disease. Ann. Hepatol. 2018, 17, 547–560.

[39] Ye, Y.; Chen, W.; Gu, M.; Xian, G.; Pan, B.; Zheng, L.; Zhang, Z.; Sheng, P. Serum globulin and albumin to globulin ratio as potential diagnostic biomarkers for periprosthetic joint infection: A retrospective review. J. Orthop. Surg. Res. 2020, 15, 1–7.

[40] Indian liver Patient Records. available online: https://github.com/aashitaarora/Classification-LiverDiseaseDataset-/blob/master/submission1_test.csv

[41Maldonado, S.; López, J.; Vairetti, C. An alternative SMOTE oversampling strategy for high-dimensional datasets. Appl. Soft Comput. 2019, 76, 380–389.

[42] Dritsas, E.; Fazakis, N.; Kocsis, O.; Moustakas, K.; Fakotakis, N. Optimal Team Pairing of Elder Office Employees with Machine Learning on Synthetic Data. In Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), Chania Crete, Greece, 12–14 July 2021; pp. 1–4.

[43] Jain, D.; Singh, V. Attribute selection and classification systems for chronic disease prediction: A review. Egypt. Inform. J. 2018, 19, 179–189.

[44] Nusinovici, S.; Tham, Y.C.; Yan, M.Y.C.; Ting, D.S.W.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.Y. Logistic regression was as good as machine learning for

predicting major chronic diseases. J. Clin. Epidemiol. 2020, 122, 56–69.

[45]. Ghosh, S.; Dasgupta, A.; Swetapadma, A. A study on support vector machine based linear and non-linear pattern classification. In Proceedings of the 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 21–22 February 2019; pp. 24–28.

[46]. Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." International Journal of Data Mining & Knowledge Management Process 8.2 (2018): 01-09.

[47] Cunningham, P.; Delany, S.J. k-Nearest neighbour classifiers-A Tutorial. ACM Comput. Surv. (CSUR) 2021, 54, 1–25.

[48] Dong, X.; Yu, Z.; Cao,W.; Shi, Y.; Ma, Q. A survey on ensemble learning. Front. Comput. Sci. 2020, 14, 241–258.

[49] Palimkar, P.; Shaw, R.N.; Ghosh, A. Machine learning technique to prognosis diabetes disease: RRandom Forest classifier approach. In Advanced Computing and Intelligent Technologies; Springer: Berlin/Heidelberg, Germany, 2022; pp. 219–244.

[50] González, S.; García, S.; Del Ser, J.; Rokach, L.; Herrera, F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives, and opportunities. Inf. Fusion 2020, 64, 205–237.

[51] Handelman, G.S.; Kok, H.K.; Chandra, R.V.; Razavi, A.H.; Huang, S.; Brooks, M.; Lee, M.J.; Asadi, H. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. Am. J. Roentgenol. 2019, 212, 38–43.

[52] Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021, 10, 593.

[53] Swapna, K.; Prasad Babu, M. Critical analysis of Indian liver patient's dataset using ANOVA method. Int. J. Eng. Technol 2017,7, 19–33.

[54] Gulia, A.; Vohra, R.; Rani, P. Liver patient classification using intelligent techniques. Int. J. Comput. Sci. Inf. Technol. 2014,5, 5110–5115.

[55] Kumar, P.; Thakur, R.S. Early detection of the liver disorder from imbalance liver function test datasets. Int. J. Innov. Technol.Explor. Eng. 2019, 8, 179–186.

[56]. Jin, H.; Kim, S.; Kim, J. Decision factors on effective liver patient data prediction. Int. J. Bio-Sci. Bio-Technol. 2014, 6, 167–178.

[57] Rahman, A.S.; Shamrat, F.J.M.; Tasnim, Z.; Roy, J.; Hossain, S.A. A comparative study on liver disease prediction using supervised machine learning algorithms. Int. J. Sci. Technol. Res. 2019, 8, 419–422.

[58] M. Abdar, N. Y. Yen, and J. C.-S. Hung, "Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees," Journal of Medical and Biological Engineering, pp. 1-13, 2017.

[59] Geetha, C.; Arunachalam, A. Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India,27–29 January 2021; pp. 1–4.

[60] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[61] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002, 16:341–378.

[62] Dritsas E, Trigka M. Supervised Machine Learning Models for Liver Disease Risk Prediction. *Computers*. 2023;12(1):19. https://doi.org/10.3390/computers12010019.

[63] I. Hanif and M. M. Khan, "Liver Cirrhosis Prediction using Machine Learning Approaches," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 2022, pp. 0028-0034, doi: 10.1109/UEMCON54665.2022.9965718.

[64] Sachdeva, R.K., Bathla, P., Rani, P. et al. A systematic method for diagnosis of hepatitis disease using machine learning. Innovations Syst Softw Eng 19, 71–80 (2023). https://doi.org/10.1007/s11334-022-00509-8.

[65] H. S. Yadav and R. K. Singhal, "Classification and Prediction of Liver Disease Diagnosis Using Machine Learning Algorithms," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp.1-6, doi: 10.1109/INOCON57975.2023.10101221.