P-ISSN : 2715-2448 | E-ISSN : 2715-7199
Vol.4 No.2 July 2023
Buana Information Technology and Computer Sciences (BIT and CS)

Liver Disease Prediction Model Based on Oversampling Dataset with RFE Feature Selection using ANN and AdaBoost algorithms

Ahmed Sami Jaddoa[1], Samah J. Saba[2], Elaf A.Abd Al-Kareem[3]
[1] Business Informatics College, University of Information Technology and Communications, Iraq
[2] Department of Computer science, Science of College, University of Diyala, Iraq
[3] Department of Sharia, College of Islamic Sciences, University of Diyala, Iraq
ahmed.sami@uoitc.edu.iq [1], Samah.j.saba@gmail.com [2], elaaf.ali1989@gmail.com [3]

Abstract
Liver disease counts are one of the most prevalent diseases all over the world and they are becoming very common these days and can be dangerous. Liver diseases are increasing all over the world due to different factors such as excess alcohol consumption, drinking contaminated water, eating contaminated food, and exposure to polluted air. The liver is involved in many functions related to the human body and if not functioned properly can affect the other parts too. Predication of the disease at an earlier stage can help reduce the risk of severity. This paper implemented oversampling dataset, feature selecting attributes, and performance analysis for the improvement of the accuracy of classification of liver patients in 3 phases. In the first phase, the z-score normalization algorithm has been implemented to the original liver patient data-sets that has been collected from the UCI repository and then works on oversampling the balanced dataset. In the second phase, feature selection of attributes is more important by using RFE feature selection. In the third phase, classification algorithms are applied to the data-set. Finally, evaluation has been performed based upon the values of accuracy. Thus, outputs shown from proposed classification implementations indicate that ANN algorithm performs better than AdaBoost algorithm with the help of feature selection with a 92.77% accuracy.

Keywords: Machine learning, Classification, Feature selection, RFE, ANN, AdaBoost, and Liver.

Abstrak
Hitungan penyakit hati adalah salah satu penyakit yang paling umum di seluruh dunia dan menjadi sangat umum akhir-akhir ini dan bisa berbahaya. Penyakit hati meningkat di seluruh dunia karena berbagai faktor seperti konsumsi alkohol berlebihan, minum air yang terkontaminasi, makan makanan yang terkontaminasi, dan paparan udara yang tercemar. Hati terlibat dalam banyak fungsi yang berkaitan dengan tubuh manusia dan jika tidak berfungsi dengan baik dapat mempengaruhi bagian lain juga. Predikasi penyakit pada tahap awal dapat membantu mengurangi risiko keparahan. Makalah ini mengimplementasikan dataset oversampling, atribut pemilihan fitur, dan analisis kinerja untuk peningkatan akurasi klasifikasi pasien hati dalam 3 fase. Pada tahap pertama, algoritme normalisasi z-score telah diimplementasikan ke kumpulan data pasien hati asli yang telah dikumpulkan dari repositori UCI dan kemudian bekerja pada oversampling kumpulan data yang seimbang. Pada tahap kedua, pemilihan fitur atribut lebih penting dengan menggunakan pemilihan fitur RFE. Pada fase ketiga,

algoritma klasifikasi diterapkan pada kumpulan data. Akhirnya, evaluasi telah dilakukan berdasarkan nilai-nilai akurasi. Dengan demikian, keluaran yang ditunjukkan dari implementasi klasifikasi yang diusulkan menunjukkan bahwa algoritma JST memiliki kinerja yang lebih baik daripada algoritma AdaBoost dengan bantuan pemilihan fitur dengan akurasi 92,77%.

Kata kunci: Pembelajaran mesin, Klasifikasi, Pemilihan fitur, RFE, ANN, AdaBoost, dan Liver

I. INTRODUCTION

Liver disease can be defined as liver inflammation that results from the actions of bacteria, or toxic materials so that liver doesn’t properly operate anymore. According to the reports that have been conducted by World Health Organization (WHO) 2005 there has been an estimate that 7.6 million patients had died from cancer and 84 million individuals would die over the next decade. This data had shown that the liver cancer represents 6th most widespread cancer type worldwide and it is the 3rd-largest death cause along with the development. It’s unavoidable that technology development and easier access to internet have made it easier to identify liver disease and become big supporters of dealing with special need illnesses [1]. Machine Learning (ML) represents an Artificial Intelligence (AI) part that allows the system to get knowledge without any explicit knowledge. The supervised algorithms take advantage of the human inputs and outputs for prediction accuracy and training process, which is why, they are utilized for a variety of the applications of classification. Thus, ML application had extended to the health-care also. A very significant problem in the health-care is the rising numbers of the liver disease patients. Liver is one of the most vital organs with some functionalities such as detoxification of chemicals, bile production, and productions of vital protein types for the blood clotting [2].Feature selection has also been referred to as the Instance Selection, Attribute Selection, Variable Selection, Data Selection, Feature Construction, or Feature Extraction. It is utilized for the data reduction by redundant and removing irrelevant data for increasing data mining accuracy. Feature Selection chooses many relevant features from original features [3].Classification has been defined as one of the crucial tasks in DM and ML, due to the fact that it is aimed at categorizing every instance in the dataset to distinctive groups on the basis of information that has been identified by its features. In addition to that, a major DM task is the data classification. it has been attempted to create classifier identifying diabetes at minimal cost and with optimal performance [4][5].

II. LITERATURE REVIEW

Over the recent years, various researches have been performed to classify liver patients. S. Jain et al. [6], proposed a paper based on the Indian Liver Patient Dataset that has a variety of symptoms for around 600 patients. this work is aimed at the evaluation of several Intelligent Technique outputs, such as K-NN, XGBoost, support vector machines (SVM), and decision tree with the ratio of the training set to testing set being 80% and 20% respectively. And results have shown that K-NN gives an accuracy of 64%, the SVM model gives a 66% accuracy, the Decision Tree model gives an 81% accuracy, and XGBoost gives an accuracy of 91%.

G. Jamila et al. [7], The proposed model for the prediction of liver cirrhosis sickness employed Naive Bayesian, Classification and Regression Tree (CART), and SVMs with 10-fold cross-validation. Accuracy, recall, precision, and F1 score were used for the evaluations of the model's performance.

Among all the strategies used in this study, SVM technique produces the optimal results, with an accuracy of 73%, precision of 73%, recall of 100%, and F1 score of 84%.

G. S. Harshpreet Kaur [8], This study has been based upon the prediction of the liver diseases with the use of ML algorithms. The prediction of the liver diseases involves many different levels of steps, such as: preprocessing, classification and feature extraction. In this paper, a hybrid classification approach has been suggested for the prediction of liver diseases, and Data-sets have been collected from Kaggle data-base of Indian liver patient records. The suggested model was able to achieve a 77.58% accuracy.

M. Ghosh et al. [9], aimed at evaluating a number of the ML outputs, such as random forest, logistic regression, XGBoost, SVMs, AdaBoost, decision tree and K-NN for prediction and diagnosis of the chronic liver disease. The algorithms of the classification have been assessed on the basis of different criteria of measurement, like the accuracy, F1 score, precision, recall, area under the curve (AUC), and specificity. Amongst algorithms, random forest exhibited superior performance in the prediction of liver diseases with 83.7% accuracy.

N. Nahar et al. [10], analyzed a new and efficient method of ensemble learning for classification of liver diseases, where 5 ensemble algorithms, namely AdaBoost, BeggRep, LogitBoost, Begg-J48, and random Forest have been implemented and compared based on accuracy, FPR, RMSE TPR, and ROC curve. LogitBoost outperformed the rest of the ensemble methods, where its accuracy has been 71.53%. This paper has codified an effective process for diagnosing liver disease using deep learning giving it a web-based approach. The model attained an accuracy of 67.6 percent and this model predicts whether the user is having a liver disease or not.

III. MATERIALS AND METHOD

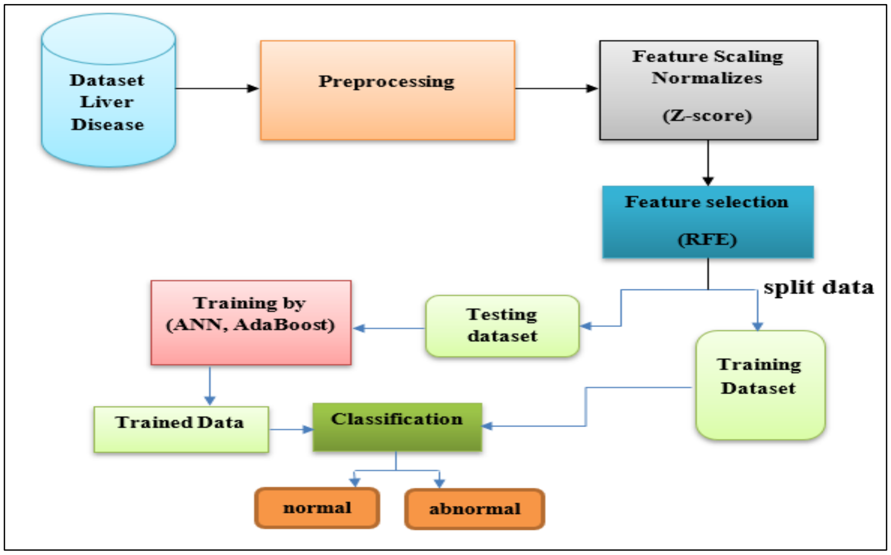

Fig.1. Overall Process of Liver Disease Model

A. Dataset and Attributes

Presently, there is a wide range of the data-sets related to liver diseases. In the present paper, ILPD has been utilized, it includes 583 rows and 2 classes. Where 1st class is associated with the patient records (PRs) of the liver disease and includes 416 records, the 2nd one is for the non-liver (PR) and consists of 167 records determined with the use of the summation of every sector field. Fig1 illustrates the distribution of the data in data-set. In general, the data-set includes 11 columns for 142 females and 441 male patients. Details have been listed in Table1.

Table1. Attributes of the Dataset

No	Attributes	Type	Range
1	Age: Patient Age	Interval	[4-90]
2	Gender: Patient Gender	Nominal	[Female- Male]
3	TB: Total Bilirubin	Interval	[0.40-75]
4	DB: Direct Bilirubin	Interval	[0.10-19.70]
5	Alkphos: Alkaline Phosphotase	Interval	[63-2,110]
6	Sgpt Alamine: Amino-transferase	Interval	[10-2,000]
7	Sgot Aspartate: Amino-transferase	Interval	[10-4,929]
8	TP: Total Protiens	Interval	[2.70-9.60]
9	ALB: Albumin	Interval	[0.90-5.50]
10	A/G Ratio: Ratio of Albumin and Globulin	Interval	[0.30-2.80]
11	Selector field *	Binary	[1-2]

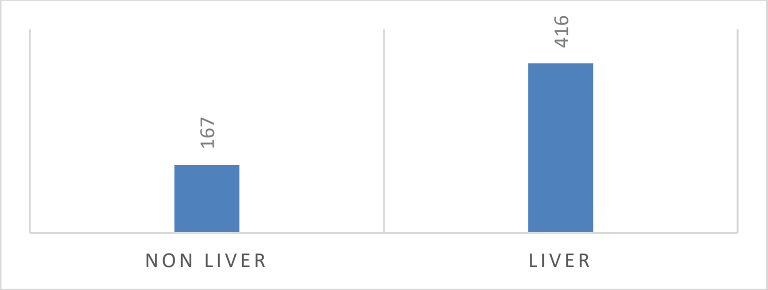

Fig. 1. The number of patients in the dataset

B. Dataset Pre-Processing

Pre-processing can be defined as a highly vital stage in ML classification as the cleaner the data, then the better are the result of classification tends to be [11]. The methods of preprocessing that have been applied in the model can be explained as:

a. Reducing noisy data: There are 2 data noise types in ML, which include: class noise and attribute noise. None-the-less, for the maximum accuracy in suggested model, the attribute noise is decreased for enhanced accuracy with the use of panda library.

b. Data transformation: which indicates the process of the reorganization or re-structuring of the raw data. It's utilized for the purpose of transforming the raw data to proper format allowing the data mining to obtain the strategic information faster and in a more effective way.

c. Standard scalar: which transforms the data in a way that its distribution has an average value of 0 as well as a standard deviation that equals to 1. The aggregate functions conduct the operations on column values them return one value.

C. Oversampling

Oversampling refers to the random duplication of the minority class values. As we have already seen, the IPLD dataset has 167 non-liver samples and 416 liver samples. Therefore, it may suffer from imbalanced class distribution issue that the class of the majority may bias prediction. To overcome this problem, Random Oversampling is used to increase the majority of class samples.

D. Feature Scaling

Feature Scaling normalizes feature values in a pre-defined range. It's a very vital step for building a machine learning model. It reduces the training time and sometimes helps to achieve faster convergence for many machines learning. Scaling using mean and standard deviation may suffer if a dataset contains too many outliers. We have used the Z-score outlier detection technique to detect the outliers and handle those outliers using Robust Scaling.

E. Feature Selection

Feature selection can be defined as the process of the selection of significant characteristics strongly associated with output from data-set for faster model training, decreased dimensionality, reduced complexity, improved accuracy and straightforward interpretation. Significant bio-markers/variables have been obtained from records of clinical information and lab tests of the patients with the use of the ML and statistical data mining algorithms. The abovementioned preprocessing tools include packages allowing feature selection [12].

Recursive Feature Elimination (RFE)

RFE can be defined as feature selection approach of a wrapper type. Internally, it utilizes filter-based approaches; none-the-less, it differs from filter method. It has 2 significant options of configuration, which include: i. it determines the number of the features that are to be chosen, ii. it sets ML algorithm in the feature selection. In initial case, it searches a sub-set of the features through the consideration of all of the features that are present in training data-set and eliminates features until the needed number of the features is left. In 2nd case, it utilizes an ML algorithm and ranks characteristics based on their significance. It discards least significant features then repeats model fitting steps. The entire process is repeated to the point where the stated number of the features is left [13].

F. Dataset Splitting

The data-set is split to data for process analysis training and testing. In that, 80% of the data has been utilized for the training and 20% of it has been used for the testing.

G. Classification Techniques

Classification is a model used to predict the future behavior of the data by classifying the records into predefined classes. In the classification, precise disease detection with the use of the testing and training dataset [14]. It proposed 2 ML models for building prediction. Initially, the training data has been trained across 2 ML models, such as Neural Network and AdaBoost are predicted based on a trained model of learning, one by one, and after that, test the data. Some of the parameters that include the precision, accuracy, and recall are finally compared with some algorithms that have been explained above.

a. Artificial Neural Network Classifier

An ANN [15] is a simulation of the working of the biological neural networks. Each one of the nodes has been modeled after a neuron, which is why, it is referred to as artificial neurons as well. An NN is made up of several layers, every one of which has a number of the nodes. The typical NN has been represented by Fig2.

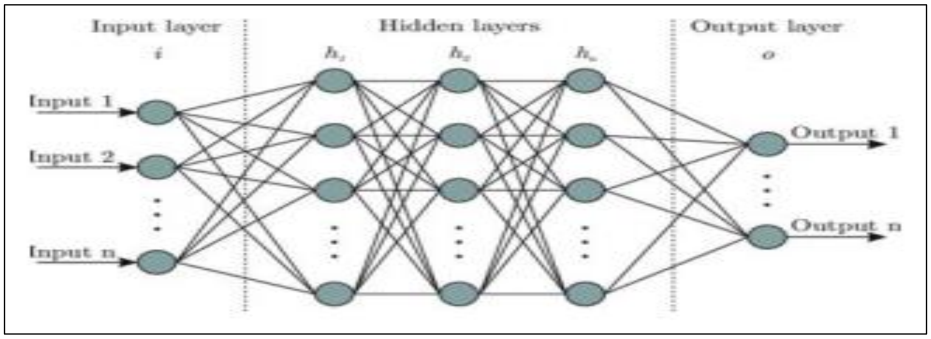

Fig2. Diagram of a typical ANN

Basically, there are 3 components in the typical ANN:

- **Input Layer –** one layer whose number of the nodes is dependent upon the input dimensions. The input layer applies a transform to NN's input and passes that along as input to hidden layers.
- **Output Layer -** which is the last layer of an NN, the dimensions of which have been characterized by the output. This layer conducts a functionality on hidden layer's output prior to the production of the results.
- **Hidden Layer -** Those layers represent the algorithm's crux. They conduct all of the calculations on input for the purpose of producing output. The work of those layers is not known. Which is why, only weights and parameters that have been provided to those layers may be tweaked for the purpose of producing the needed results. A network becomes deeper with the increase of the number of the hidden layers. Each one of the nodes in a network is referred to as a perceptron, which has been depicted in Fig3. A perceptron is made up of 2 parts, which are: a sum of inputs and activation function on summation. A certain node takes weighted summation of its inputs then passes it to linear or nonlinear activation function.

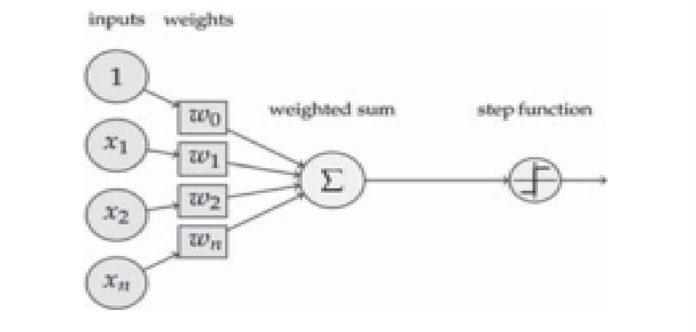

Fig3. A Diagram of the Perceptron

The equation for certain perceptron has been depicted by 1. Weighted summation of inputs (x.w) is passed through activation function (f) besides bias value (b). It may be denoted as product of vector dot, where n represents the number of the inputs for each node. The activation function produces output prediction that has been provided as set of the inputs. Bias term has been added to computation for the purpose of helping in the enhancement of the learning of the perceptron.

$$z = f\ (b + x.w) = f\ (b + \textstyle\sum_{i=1}^{n} x_i\, w_i) \qquad (1)$$

each perceptron utilizes step function as the activation function. In a set of the perceptron's, which is ANN (referred to as the Multi-Layer Perceptron as well), each one of the layers may have a separate activation function.

b. AdaBoost Algorithm

AdaBoost algorithm includes the use of very short (1-level) decision trees as weak learners added in a sequential manner to the set. Every one of the consequent models tries correcting predictions that have been made by the model before it in a sequence. It combines several of the average or weak predictors for the purpose of building strong predictor [16].

H. Performance measure

Performance measure of different machine learning algorithms is analyzed by considering measures such as [17].

- **Confusion Matrix -** The confusion Matrix is a table used in performance measures that helps in easy visualization as well as in distinguishing true positives, true negatives, false positives and false negatives.
- **Accuracy -** Accuracy measure is calculated by considering the ratio of the observations that have been correctly predicted to total number of the observations.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

- **Precision -** It represents the percentage of true positives out of all the predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (3)$$

- **Sensitivity -** Out of the total positive, what percentage are predicted positive.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (4)$$

- **Specificity** – it represents True negative rate which is the proportion of the negative tuples which have been identified correctly.

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (5)$$

IV. RESULTS AND DISCUSSION

On the implementation of algorithms that have been mentioned in previous section, the following results have been obtained:

Table1. Confusion matrix

Actual / Predicted	Normal	Abnormal
Normal	TP	FN
Abnormal	FP	TN

Table 2. confusion matrix of **ANN** without **RFE** Feature selection

Actual / Predicted	Normal	Abnormal
Normal	15	3
Abnormal	5	60

Table 3. confusion matrix of **ANN** with **RFE** Feature selection

Actual / Predicted	Normal	Abnormal
Normal	16	1
Abnormal	5	61

Table 4. Performance measure of **ANN** model without and with **RFE** feature selection

Model	Accuracy	Precision	Sensitivity	Specificity
ANN without	90.36%	75%	83.3%	92.3%

RFE				
ANN with RFE	92.77%	76.1%	94.1%	92.4%

Table 5. confusion matrix of **AdaBoost** without **RFE** Feature selection

Actual / Predicted	**Normal**	**Abnormal**
Normal	12	5
Abnormal	6	60

Table 6. confusion matrix of **AdaBoost** with **RFE** Feature selection

Actual / Predicted	**Normal**	**Abnormal**
Normal	13	4
Abnormal	5	61

Table 7. Performance measure of **AdaBoost** model without and with **RFE** feature selection

Model	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
AdaBoost without RFE	86.74%	66.6%	70.5%	90.9%
AdaBoost with RFE	89.15%	72.2%	76.4%	92.4%

Fig. 4. Show accuracy of **ANN** and **AdaBoost** models without and with **RFE** feature selection

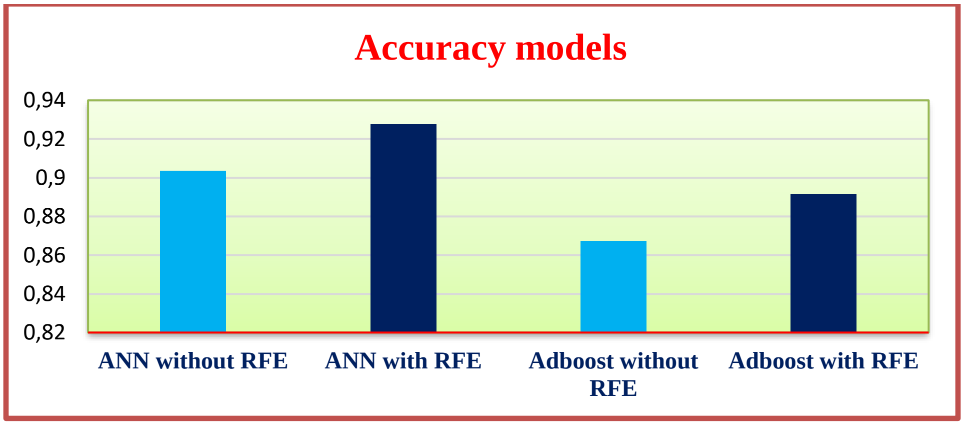

Fig. 4. Accuracy of ANN and AdaBoost models

V. CONCLUSIONS

This work presented a model for prediction of liver disease occurrence probability. The analyses and evaluations of suggested model have shown that it's highly sufficient and easy to utilize and implement. Two ML algorithms have been applied to ILPD data-set for classified liver patients. In the data preprocessing issue of imbalanced class distribution, an oversampling technique (Random Over Sampling) is used, and used the Z-score outlier detection technique to detect the outliers and handle those outliers using Robust Scaling. Then applied RFE feature selection specifies the number of characteristics to be chosen is used for achieving better performance and for achieving an enhanced result, we have applied ANN and AdaBoost algorithms. From the analysis of experimental results, the ANN algorithm has achieved the highest accuracy of 92.77%.

References

[1] H. Hartatik, M. B. Tamam, and A. Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," *2020 2nd Int. Conf. Cybern. Intell. Syst. ICORIS 2020*, pp. 1–5, 2020, doi: 10.1109/ICORIS50180.2020.9320797.

[2] M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 778–782, 2020, doi: 10.1109/ICACCS48705.2020.9074368.

[3] M. A. Khadija and N. A. Setiawan, "Detecting Liver Disease Diagnosis by Combining SMOTE, Information Gain Attribute Evaluation, and Ranker," *ITSMART J. Teknol. dan Inf.*, vol. 9, no. 1, pp. 13–17, 2020.

[4] A. S. Jaddoa, Z. Tariq, and M. Al-ta, "COMPARISON OF DATA MINING ALGORITHMS FOR DIAGNOSIS OF DIABETES MELLITUS," vol. 10, no. 2, pp. 1–8, 2021.

[5] R. Ahmed, S. Jaddoa, P. Ziyad, and T. Mustafa, "Diagnosis of Diabetes Mellitus using Hybrid Techniques for Feature Selection and Classification," pp. 1650–1663, 2021.

[6] S. Jain, R. Sharma, and R. Rajkamal, "EasyChair Preprint Classification of Liver Diseases Using Intelligent Techniques Classification of Liver Diseases Using Intelligent Techniques," 2021.

[7] G. Jamila, G. M. Wajiga, Y. M. Malgwi, and A. H. Maidabara, "A Diagnostic Model for the Pediction of Liver Cirrhosis using Machine Learning Teachniques," *Comput. Sci. IT Res. J.*, vol. 3, no. 1, pp. 36–51, 2022, doi: 10.51594/csitrj.v3i1.296.

[8] G. S. Harshpreet Kaur, "The Diagnosis of Chronic Liver Disease using Machine Learning Techniques," *Inf. Technol. Ind.*, vol. 9, no. 2, pp. 554–564, 2021, doi: 10.17762/itii.v9i2.382.

[9] M. Ghosh *et al.*, "A comparative analysis of machine learning algorithms to predict liver disease," *Intell. Autom. Soft Comput.*, vol. 30, no. 3, pp. 917–928, 2021, doi: 10.32604/iasc.2021.017989.

[10] N. Nahar, F. Ara, M. A. I. Neloy, V. Barua, M. S. Hossain, and K. Andersson, "A Comparative Analysis of the Ensemble Method for Liver Disease Prediction," *ICIET 2019 - 2nd Int. Conf. Innov. Eng. Technol.*, pp. 23–24, 2019, doi: 10.1109/ICIET48527.2019.9290507.

[11] S. Afrin *et al.*, "Supervised machine learning based liver disease prediction approach with LASSO feature selection," *Bull. Electr. Eng. Informatics*, vol. 10, no. 6, pp. 3369–3376, 2021, doi: 10.11591/eei.v10i6.3242.

[12] N. Tanwar and K. F. Rahman, "Machine learning in liver disease diagnosis: Current progress and future opportunities," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012029.

[13] R. C. Poonia *et al.*, "Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease," *Healthc.*, vol. 10, no. 2, 2022, doi: 10.3390/healthcare10020371.

[14] S. Kefelegn, "Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey," vol. 118, no. 9, pp. 765–770, 2017, [Online]. Available: http://www.ijpam.eu.

[15] S. Gupta, G. Karanth, N. Pentapati, and V. R. B. Prasad, "A Web Based Framework for Liver Disease Diagnosis using Combined Machine Learning Models," *Proc. - Int. Conf. Smart Electron. Commun. ICOSEC 2020*, no. Icosec, pp. 421–428, 2020, doi: 10.1109/ICOSEC49089.2020.9215454.

[16] A. Khatavkar, P. Potpose, and P. Pandey, "Smart Health Prediction System," vol. 5, no. 02, pp. 1550–1552, 2017.

[17] B. K. Mengiste, H. K. Tripathy, and J. K. Rout, "Analysis and Prediction of Cardiovascular Disease Using Machine Learning Techniques," *Lect. Notes Electr. Eng.*, vol. 708, no. 2, pp. 133–141, 2021, doi: 10.1007/978-981-15-8685-9_13.

.