VOL. 18, NO. 17, SEPTEMBER 2023 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PREDICTION OF LIVER DISEASES WITH RANDOM FOREST CLASSIFIER WITH PRINCIPAL COMPONENT FEATURE EXTRACTION

M. Rekha Sundari[1], P. Raga Manognya[2], Rishi V.[3], A. Venkata Mahesh[4] and Mugada Swetha[5]
[1]Department of Information Technology, ANITS Engineering College, Andhra Pradesh, India
[2]Department of Computer Science Engineering, Gayatri Vidhya Parishad College of Engineering for Women, Visakhapatnam, India
[3]Computer Science and Engineering, GITAM (Deemed to be University), Andhra Pradesh, India
[4]Department of Computer Science and Engineering, GIET UNIVERSITY, Gunupur, Orissa, India
[5]Department of Computer Science and Engineering, Vignans Institute of Engineering for Women, Kapujagarajupeta, Visakhapatnam, India
E-Mail: mrekhasundari.it@anits.edu.in

ABSTRACT

Liver disease is one of the most major illnesses that has a negative influence on the normal, healthy stature of a human being. This is because many different factors may lead to liver disease. The liver is the organ that is most often afflicted by liver disease even though it is the largest organ found inside the body. A few examples of the various subtypes of liver illness include fatty liver disease, cirrhosis, hepatitis, chronic liver disease, liver cancer, liver tumors, and other kinds of liver disease. This study creates machine learning algorithms to enhance the prediction of liver disease using various data balancing techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) with Edited Nearest Neighbourhood (ENN). The traits that were derived from SMOTE-ENN balanced features are then normalised using the principal component analysis (PCA) method. In addition, approaches like as correlation and skewness are used to clean the dataset and minimize the number of features. Finally, the prediction operation is carried out by a Random Forest classifier with PCA extracted and SEMOTE-ENN balanced features. The simulations conducted liver-disease dataset show that the proposed method results in improved performance over existing methods.

Keywords: liver disease, synthetic minority oversampling technique, encoded nominal and continuous, edited nearest neighbourhood, random forest.

Manuscript Received 30 June 2023; Revised 24 October 2023; Published 8 November 2023

1. INTRODUCTION

Before commencing the process of model creation, it is vital to have dependable data collection that adequately depicts the different sickness types [1]. It is challenging to recognize liver disorders in their early stages because the liver may continue to operate normally while being damaged. If liver disease is diagnosed at an earlier stage, the patient has a better chance of recovery. Liver failure is more common among Indians. According to predictions, India would become the world's hub for liver illnesses by 2025 [2]. The high prevalence of liver infections in India may be attributed to several reasons, including a lack of physical activity, rising alcohol use, and smoking. There are more than a hundred different kinds of liver infections that have been found [3]. Consequently, the medical community will benefit greatly from the creation of a tool that may aid in the identification of disorders. The length of time patients must wait to see liver specialists like endocrinologists will be reduced by automatic categorization tools for liver problems (likely mobile or web enabled), and these technologies will help doctors make accurate assessments of their patients [4]. The procedure of collecting samples from patients for use in medical diagnosis may be time-consuming and costly. Obtaining all the data necessary for a more precise diagnosis usually necessitates doing many tests on the patient or collecting many samples from them. Urinalysis, complete blood count (CBC) [5], and comprehensive metabolic panels are the diagnostic tests that are used most often (CMP). The cost of these evaluations is often lower than that of others, even though they could still give useful information. The liver is responsible for a wide variety of metabolic processes, including the creation and storage of glucose, the removal of toxins [6], the manufacture of digestive enzymes, the control of erythrocytes, and protein synthesis. Cirrhosis, fibrosis, and chronic hepatitis are all examples of ailments that are considered chronic and may have a significant impact on quality of life. There are several potential causes of hepatitis, including viral infections (such as hepatitis C virus) and autoimmune reactions. Inflammation generated by a hepatitis infection may damage liver tissues and lead to scarring [7]. The medical word for mild scarring is fibrosis, whereas the term for severe scarring and liver damage is cirrhosis. Liver fibrosis and cirrhosis may also be caused by alcoholism and non-alcoholic fatty liver disease [8]. By detecting liver disease early, before it becomes irreversible cirrhosis, it is feasible to prevent the condition from progressing to the point where the liver fails. Liver illness may sometimes be diagnosed with the use of diagnostic tests like a cholecystogram or a biopsy [9]. Diagnosis and treatment of a liver problem include

tracing its symptoms back to their sources. Aminotransferases, alpha-aminotransferases, and aminotransferase ALT all play crucial roles in gluconeogenesis [10]. These enzymes are responsible for catalyzing the process that converts ketoglutaric acid into a form of alpha-amino acid. As a protein that is not exclusive to the liver, AST may be found in many other organs and sometimes serves as a marker for illnesses that have nothing to do with the liver.

2. LITERATURE SURVEY

Classification techniques [11] are widely used across a broad range of automated medical diagnostic equipment types. Early diagnosis of liver disorders may be challenging since the liver may continue to function normally even when it is partially damaged. If liver diseases are detected and treated early on, patients have a greater chance of survival. Enzyme blood testing is one possible way to diagnose liver disease [12]. In addition, modern mobile devices are increasingly being utilized to track people's health and wellness. Additionally, automatic categorization techniques are necessary for this goal. To reduce the number of patients waiting to visit liver specialists like endocrinologists, automated classification systems for liver disorders may be useful [13]. You'll probably be able to use these resources on the go using your smartphone or the web.

The SVM classifier produces the greatest predicted performance when applied to chemical datasets, according to authors in [14]. Using the CDC Chronic Fatigue Syndrome dataset [15], Lung-Cheng Huang discovered that the nave Bayesian classifier performed better than SVM and C 4.5. based on the findings of a study that was done. The characteristics of the dataset that is going to be studied will determine which algorithm will have the highest performance. In [16] authors utilized transcriptome analysis methods, especially gene array analysis, to learn more about human liver disease by focusing on studies that used these methods. They have also presented an overview of the existing knowledge on the diversity of the transcriptome in healthy liver tissue and an explanation of the relationship between transcript and protein expressions [17]. Methods like as subtractive hybridization, differential display, and serial examination of gene expression are discussed. The major focus of this article is on the use of in-vitro cell culture methods in the study of primate entire organs. The analysis of the transcriptome of human liver disease offers a wealth of opportunities for furthering our knowledge [18] of hepatic mRNA expression in health and illness. They conclude that our understanding of human liver pathobiology has benefited greatly from prior research into the hepatic transcriptome.

In [19] authors identified the most common sickness in Taiwan, liver disease is notoriously hard to identify in its early stages. Due to its status as a leading killer, prompt identification of liver disease is of critical importance. As a result, creating a reliable diagnostic paradigm is one of the biggest obstacles in the delivery of care for liver illness. Because of this, the study reported here employs categorization and regression tree (CART) [20] and case-based reasoning (CBR) techniques to create a smart diagnostic model. To better diagnose liver illness, this project aims to develop a complete analytical platform. The goal of this study is to improve the precision with which liver illness may be identified by developing a complete analytical framework.

To evaluate the efficacy of their classification system, in [21] authors put it to a dataset including information on patients with liver illness and calculated its accuracy, precision, sensitivity, and specificity. Overall, the AP Liver Dataset outperforms the UCLA Liver Datasets in terms of accuracy, precision, sensitivity, and specificity. This is because the AP liver dataset has more informative features than the UCLA liver dataset [22], which accounts for the AP liver dataset's superior performance. Total bilirubin, direct bilirubin, indirect bilirubin, albumin, sex, age, and total proteins are all instances of such variables. In addition, there are more proteins overall in the AP liver collection. Age, sex, SGOT, SGPT, ALP, total bilirubin, direct bilirubin, total proteins [23], and albumin are all important considerations for assessing liver health. Both the AP liver data and the Taiwan liver data may have these characteristics. Regardless of the feature set utilized, KNN, Backpropagation, and SVM all show promising results on the selected dataset.

Predictions of different liver diseases were reported in [24], which used a rule-based classification approach that combines machine learning techniques. Five hundred and eighty-three patients and twelve characteristics were collected, with 441 males and 292 females making up the total. The suggested model employs a variety of data mining approaches, including the K-cross fold methodology, to provide reliable prognoses for liver illnesses. Some examples of these methods are the Support Vector Machine (SVM), Rule Induction (RI), Decision Tree (DT), Naive Bayes (NB), and Artificial Neural Network (ANN) [25]. Accuracy, sensitivity, specificity, and kappa are used to evaluate the efficacy of various data mining methods. ANOVA and Chi-square tests, two common statistical methods, are also used to investigate the dataset and feature independence in liver disease.

3. PROPOSED SYSTEM

The liver is in the right upper abdomen, close below the rib cage, which is an important organ. It is one of the body's most important organs, if not the most important. Besides helping the body get rid of potentially dangerous toxins, it also aids in keeping blood sugar levels stable. While the liver and other organs in the body can heal itself, chronic liver damage from excessive alcohol use or exposure to environmental toxins may raise the chance of liver failure. Liver transplantation, although costly and only seldom effective, is the solution. The risk

of liver failure may be reduced if liver disease is identified at an early stage. The machine-learning algorithm can estimate the likelihood of a disease based on a data set comprised of relevant health indicators from both individuals with and without the ailment.

In Figure-1, we see a block diagram depicting the proposed study methodology. The comma-separated values (.csv) format is used for this specific dataset on liver disease. Then, a pre-processing procedure is performed to remove duplicate spaces and complete blank characters. This first processing step also normalizes the whole dataset by making all the elements the same size. Following this, features are recovered from the dataset by using PCA and skewness attributes. The PCA is used here to maintain important feature quality while concurrently reducing the number of dimensions. The SMOTE-ENN balancing method is then used in the dataset to achieve class parity. Finally, the different types of prediction jobs are executed using the random forest classification process.

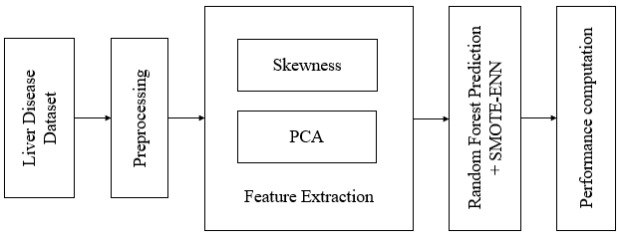

Figure-1. Block diagram of proposed system.

3.1 Pre-Processing

To utilize the raw data in a machine learning model, it must first be cleaned and formatted, a process known as "data preparation." Creating a machine learning model begins with this very first step, which is also the most important phase. Although clean and well-organized data is ideal for every machine learning effort, it is not always available. In addition, before engaging in any action that involves data, it is essential to thoroughly clean the data and arrange it appropriately. This must be done before any activity can begin. Because of this, we make use of the process of data preparation for this goal. In most cases, real-world data cannot be utilized as-is in machine learning models due to noise, missing values, and maybe an unacceptable format. This means that it is not possible to use machine learning on real-world data without some kind of transformation. Data preparation refers to the steps taken to ensure the data is clean and ready to be used in a machine-learning model. The effectiveness and precision of the machine learning model are both enhanced by this.

3.2 Feature Scaling

The final part of the machine learning "data preparation" process is termed "feature scaling." It's a technique for making the dataset's independent variables consistent within a predetermined interval. In feature scaling, we normalize all our variables such that they fall inside the same numerical range. This prevents any one variable from being given undue weight. The machine learning model will have issues if the variable is not scaled.

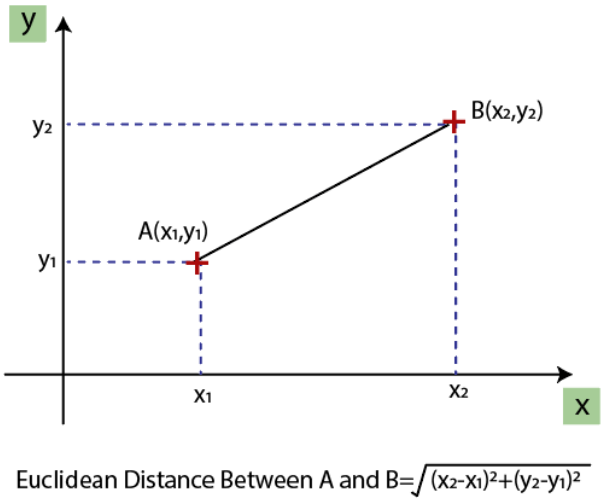

Figure-2. Feature scaling.

Since the foundation of a machine learning model is the Euclidean distance between two points, we will run into trouble if we neglect to adjust the parameters for this distance. As an example of the formula for Euclidean distance, consider as shown in Figure-2. The results of any calculation involving age and pay will be erroneous because the salary-related variables will take priority over the age-related calculations. For this reason, we need to engage in feature scaling for machine learning and put an end to the issue.

3.3 Splitting the Dataset

Our dataset is divided into a training set and a test set during the pre-processing phase of machine learning. One of the primary objectives of the data preparation operations, which makes this step significant, is the overall performance of our machine learning model. Think about the following: After training it on one dataset, we use a separate one to put our machine learning model to the test. If anything similar happens, our model may have trouble comprehending how different models are interdependent. When we give our model a fresh dataset to examine, it will perform less well even if we train it to absolute perfection and ensure that its training accuracy is incredibly high. Therefore, it is crucial to verify that a machine learning model works adequately outside of the training set and the test dataset.

Training set: A portion of the dataset to be used to train the machine learning model, and the results have already been determined.

Test set: A portion of the dataset that is used to test the machine learning model; by making use of the test set, the model makes predictions about the output.

3.4 PCA Feature Reduction

To reduce the dimensionality of data, a well-known unsupervised learning method called the PCA may be applied. As a bonus, it reduces the quantity of data that is discarded while improving the information's readability. Finding the most important aspects of a dataset is made easier, as is visualising the data in two and three dimensions. If you need to find a set of linear combinations of variables, PCA is a helpful tool. In Figure-3, several points have been represented by being plotted on a plane that only has two dimensions. In this scenario, there are two fundamental factors at play. The dominant principal component, PC1, is responsible for the bulk of the data's variance explanation. It is liable for this by its role as the main principal component. PC2 is an additional principal component that may be thought of as being orthogonal to PC1, which stands for the first principal component.

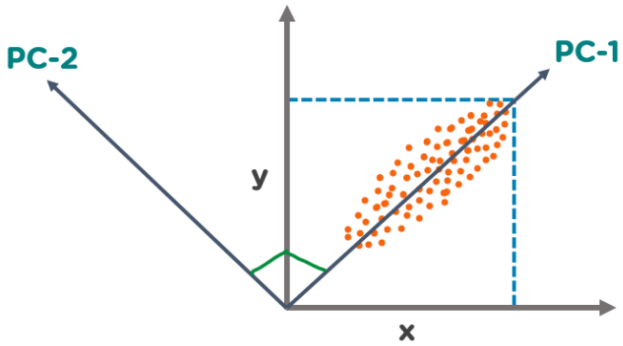

Figure-3. PCA feature analysis.

The following points show the purpose of PCA, and Table-1 shows the PCA algorithm.

- PCA is used to display data that consists of several dimensions.
- PCA is used to decrease the total number of dimensions in data about healthcare.
- PCA can assist in the process of resizing a picture.
- PCA is useful for identifying patterns in high-dimensional datasets and may be used in the field of finance to evaluate stock data and make predictions about returns.

Table-1. PCA feature extraction algorithm.

Input: SMOTE-ENN samples **Output:** PCA extracted features
The data must first be normalised. You should normalise the data before starting the PCA analysis. This will ensure that the mean value of each feature is zero and the variance value is one. Construct the covariance matrix as the second step: Create a square matrix to depict the relationship between two or more attributes in a multidimensional dataset. Discovering the Eigenvectors and Eigenvalues is the Third Step Make the appropriate computations to get the eigenvalues and eigenvectors/unit vectors. The variance matrix is obtained by multiplying the eigenvector of the covariance matrix by the eigenvalues, or eigenvalue scalars. The fourth step is to arrange the eigenvectors from most significant to least significant, and then choose the number of major components.

3.5 Random Forest Classifier

Several methods fall under the umbrella of "supervised learning," with Random Forest being one of the most popular machine learning algorithms. It may be used for analysing data for classification issues and regression in machine learning. The Random Forest algorithm is presented in Table-2.

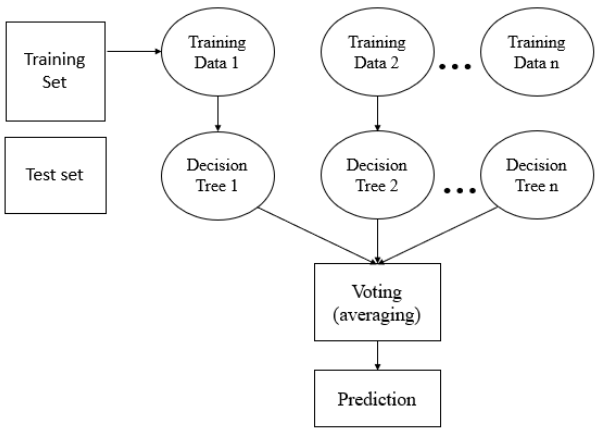

Figure-4. Random forest algorithm.

Random Forest classifier contains several decision trees on different subsets of the supplied dataset and takes the average to increase the predicted accuracy of that dataset as shown in Figure-4. To take the average to increase the forecasting accuracy of that dataset, Random Forest employs a method called ensemble learning. The random forest model does not rely on any one decision tree, but rather considers the predictions of all of the trees in the forest and then settles on the final result by giving more weight to the result from the tree with the most votes. The precision attained and the problem of overfitting may be mitigated by increasing the size of the forest.

Table-2. PCA feature extraction algorithm.

Input: PCA extracted features **Output: Classified Outcomes**
Step 1: The Random Forest technique first randomly selects n records from a dataset of size k, where k is any integer between 1 and the number of entries in the dataset. Step 2: Every decision tree is built from scratch by using the data from the individual samples. Step 3: There is an output produced by each decision tree. Step 4: Depending on whether classification or regression analysis is being carried out, the procedure for deciding the result will either include majority voting or averaging the results.

4. RESULTS AND DISCUSSIONS

This section presents a comprehensive examination of the outcomes of the simulation that was carried out utilizing the "python environment." In addition to this, the performance of the suggested approach is compared with the performance of other methods already in use using the same dataset.

4.1 Dataset

Patients from the far northeastern corner of the Indian state of Andhra Pradesh had their medical records gathered for analysis. This dataset contains data from 167 people with various liver disorders and 416 patients with various liver ailments. The "Dataset" column serves as a kind of "class label," separating the people into those with liver disease and those without (no disease). In all, this database contains medical records for 441 male patients and 142 female patients. Those patients whose ages were more than 89 were given the age "90" in the database. Columns of the dataset are:

- The patient's age.
- The patient's gender.
- The amount of total bilirubin in their blood
- The enzyme alkaline phosphatase
- Direct Bilirubin
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Proteins
- Albumin.
- the ratio of albumin to globulin

4.2 Performance Estimation

In Figure-5, the x-axis represents 1 (no disease) and 2 (disease present) and the y-axis contains several records we can see first class contains more records and the second class contains fewer records. So, it has an imbalance problem. In Figure-6, we are calculating the skewness of each attribute where negative value refers to unimportant features and positive values refer to important features. Figure-7 shows the graphical representation of skewness values. Table-3 shows the performance comparison of various methods like SVM [11], ANN [12], decision Tree [13], XGboost [14], and the proposed method. Here, the proposed method resulted in improved accuracy, precision, recall, and F1-Score metrics as compared to other approaches.

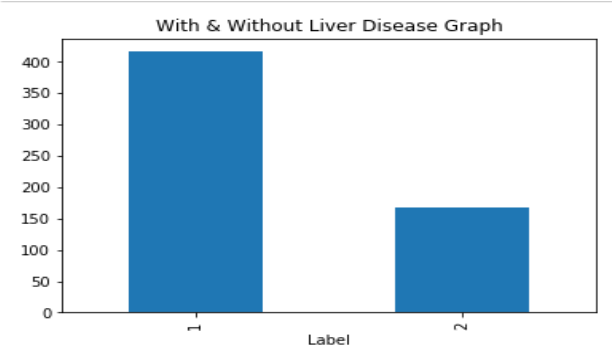

Figure-5. No of diseases.

```
Age                          -0.029385
Gender                       -1.197919
Total_Bilirubin               4.907474
Direct_Bilirubin              3.212403
Alkaline_Phosphotase          3.765106
Alamine_Aminotransferase      6.549192
Aspartate_Aminotransferase   10.546177
Total_Protiens               -0.285672
Albumin                      -0.043685
Albumin_and_Globulin_Ratio    0.805492
Label                         0.947140
dtype: float64
```

Figure-6. Skewness feature of the dataset.

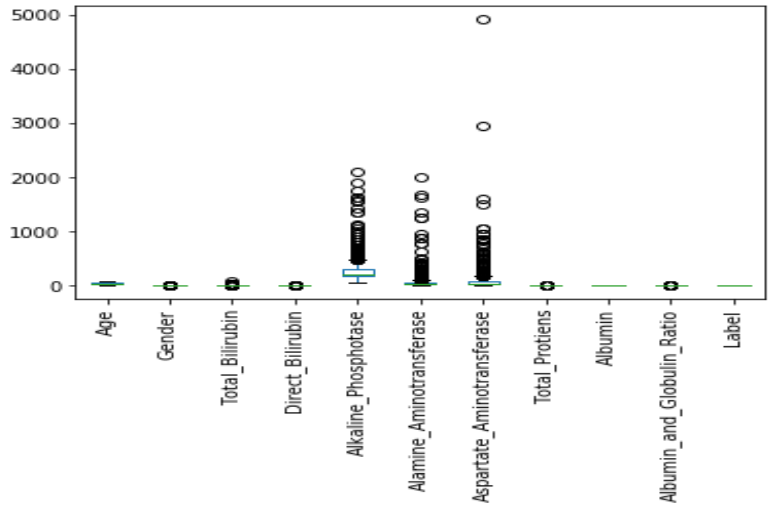

Figure-7. Graphical representation of skewness values.

Table-3. Performance comparison.

Method	Accuracy	Precision	Recall	F1-Score
SVM [11]	65.81	57.99	55.56	55.17
ANN [12]	80.83	83.87	82.36	80.75
Decision Tree [13]	76.64	76.67	76.95	76.59
XGboost [14]	84.5	85.73	84.85	84.80
Proposed Method	91.39	93.84	88.88	90.47

5. CONCLUSIONS

While the machine learning techniques discussed in this work may be useful for medical practitioners, ML classifiers should not be used in lieu of doctors' clinical judgment for determining the best course of treatment. Accurate diagnosis, inadequate data collection, expensive expenditures, and extended treatment periods are only some of the challenges that may be alleviated by using these technologies within the medical business. By enhancing the identification of risk factors and diagnostic variables, the PCA approach contributed to a reduction in the global burden of liver disease. The number of liver-related fatalities, transplants, and hospitalizations might be reduced if chronic liver disease could be detected at an earlier stage or in hidden cases employing SMOTE-ENN data balancing. If an illness can be diagnosed and treated before it progresses to a more advanced stage, the prognosis is much better. Similarly, invasive diagnostic procedures like biopsies would be used less often under these conditions. Although the emphasis of this study was on using hepatitis and chronic liver sickness characteristics for random forest training, it is reasonable to infer that the techniques may be used to distinguish between healthy individuals and other kinds of liver disease. This work can be extended with deep learning models for improved performance.

REFERENCES

[1] N. A. Shackel, D. Seth, P. S. Haber, M. D. Gorrell and G. W. McCaughan. 2006. The hepatic transcriptome in human liver disease. Comp Hepatol. 7; 5: 6. doi: 10.1186/1476-5926-5-6. PMID: 17090326; PMCID: PMC1665460.

[2] Y. Kumar and G. Sahoo. 2013. Prediction of different types of liver diseases using rule based classification model. Technol Health Care. 21(5): 417-32. doi: 10.3233/THC-130742. PMID: 23963359.

[3] S. Sontakke, J. Lohokare and R. Dani. 2017. Diagnosis of liver diseases using machine learning. 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), pp. 129-133, doi: 10.1109/ETIICT.2017.7977023.

[4] V. J. Gogi and V. M.N. 2018. Prognosis of Liver Disease: Using Machine Learning Algorithms. 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), pp. 875-879, doi: 10.1109/ICRIEECE44171.2018.9008482.

[5] J. Javad Hassannataj, S. Hamid, D. Abdollah and S. Shahaboddin. 2019. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Informatics in Medicine Unlocked, 17(100255): 2352-9148, https://doi.org/10.1016/j.imu.2019.100255.

[6] S. Ambesange, V. A, R. Uppin, S. Patil and V. Patil. 2020. Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques. 2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), pp. 98-102, doi: 10.1109/CCEM50674.2020.00030.

[7] M. A. Kuzhippallil, C. Joseph and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients", 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 778-782, doi: 10.1109/ICACCS48705.2020.9074368.

[8] S. Deshmukh, A. Lokhande, R. Wasnik and N. Singhal. 2020. Vacuole Segmentation and Quantification in Liver Images of Wistar Rat. Annu Int Conf IEEE Eng Med Biol Soc. 2020: 1396-1399. doi: 10.1109/EMBC44109.2020.9176500. PMID: 33018250.

[9] H. Hartatik, M. B. Tamam and A. Setyanto. 2020. Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms. 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), pp. 1-5, doi: 10.1109/ICORIS50180.2020.9320797.

[10] R. Kalaiselvi, K. Meena and V. Vanitha. 2021. Liver Disease Prediction Using Machine Learning Algorithms. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675756.

[11] G. Shobana and K. Umamaheswari. 2021. Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1223-1229, doi: 10.1109/ICCMC51019.2021.9418333.

[12] Straw Isabel and Honghan Wu. 2022. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Health & Care Informatics 29.1 (2022).

[13] Nam, David, et al. 2022. Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. JHEP Reports (2022): 100443.

[14] Gupta Ketan, et al. 2022. Liver Disease Prediction using Machine learning Classification Techniques. 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT). IEEE.

[15] Veeranki Sreenivasa Rao and Manish Varshney. 2022. Intelligent Techniques and Comparative Performance Analysis of Liver Disease Prediction. International Journal of Mechanical Engineering 7.1: 489-503.

[16] Liu Yang, et al. 2022. Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. Cell Metabolism 34.5: 719-730.

[17] Fujiwara Naoto, et al. 2022. Molecular signatures of long-term hepatocellular carcinoma risk in nonalcoholic fatty liver disease. Science translational medicine 14.650: eabo4474.

[18] Takahashi Satoko, et al. 2022. Prediction and validation of nonalcoholic fatty liver disease by fatty liver index in a Japanese population. Endocrine Journal. 69.4: 463-471.

[19] Vilar-Gomez, Eduardo, and Naga Chalasani. 2018. Non-invasive assessment of non-alcoholic fatty liver disease: Clinical prediction rules and blood-based biomarkers. Journal of hepatology 68.2: 305-315.

[20] Singh Ajay S., Md Irfan, and Abhishek Chowdhury. 2018. Prediction of liver disease using classification algorithms. 2018 4th international conference on computing communication and automation (ICCCA). IEEE.

[21] Azam M. S., Rahman A., Iqbal S. H. S. and Ahmed M. T. 2020. Prediction of liver diseases by using few machine learning based approaches. Aust. J. Eng. Innov. Technol. 2(5): 85-90.

[22] Srivastava A., Kumar V. V., Mahesh T. R. and Vivek V. 2022, April. Automated Prediction of Liver

Disease using Machine Learning (ML) Algorithms. In 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-4). IEEE.

[23] Rahman A. S., Shamrat F. J. M., Tasnim Z., Roy J. and Hossain, S. A. 2019. A comparative study on liver disease prediction using supervised machine learning algorithms. International Journal of Scientific & Technology Research. 8(11): 419-422.

[24] Priya, M. Banu, P. Laura Juliet, and P. R. Tamilselvi. "Performance analysis of liver disease prediction using machine learning algorithms. Int. Res. J. Eng. Technol 5.1 (2018): 206-211.

[25] Singh, Jagdeep, Sachin Bagga, and Ranjodh Kaur. 2020. Software-based prediction of liver disease with feature selection and classification techniques. Procedia Computer Science. 167: 1970-1980.