Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 6, December 2021, pp. 3369~3376
ISSN: 2302-9285, DOI: 10.11591/eei.v10i6.3242

Supervised machine learning based liver disease prediction approach with LASSO feature selection

Saima Afrin[1], F. M. Javed Mehedi Shamrat[2], Tafsirul Islam Nibir[3], Mst. Fahmida Muntasim[4], Md. Shakil Moharram[5], M. M. Imran[6], Md Abdulla[7]
[1,3,4,5,7]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[6]Cefalo Bangladesh Limited, Dhaka, Bangladesh

Article Info

Article history:

Received Jul 17, 2021
Revised Sep 18, 2021
Accepted Oct 16, 2021

Keywords:

10 fold cross-validation
Classification
LASSO
Liver disease
Machine learning
Supervised algorithms

ABSTRACT

In this contemporary era, the uses of machine learning techniques are increasing rapidly in the field of medical science for detecting various diseases such as liver disease (LD). Around the globe, a large number of people die because of this deadly disease. By diagnosing the disease in a primary stage, early treatment can be helpful to cure the patient. In this research paper, a method is proposed to diagnose the LD using supervised machine learning classification algorithms, namely logistic regression, decision tree, random forest, AdaBoost, KNN, linear discriminant analysis, gradient boosting and support vector machine (SVM). We also deployed a least absolute shrinkage and selection operator (LASSO) feature selection technique on our taken dataset to suggest the most highly correlated attributes of LD. The predictions with 10 fold cross-validation (CV) made by the algorithms are tested in terms of accuracy, sensitivity, precision and f1-score values to forecast the disease. It is observed that the decision tree algorithm has the best performance score where accuracy, precision, sensitivity and f1-score values are 94.295%, 92%, 99% and 96% respectively with the inclusion of LASSO. Furthermore, a comparison with recent studies is shown to prove the significance of the proposed system.

This is an open access article under the CC BY-SA license.

Corresponding Author:

F. M. Javed Mehedi Shamrat
Department of Software Engineering
Daffodil International University
102/1, Sukrabad, Mirpur Road, Dhaka 1207, Bangladesh
Email: javedmehedicom@gmail.com

1. INTRODUCTION

The liver is a vital organ in the human body that performs the functionalities like the production of bile, chemicals detoxification, and source of important proteins which is for blood clotting. In recent years, a significant increase of various liver diseases has been observed around the globe. In India, the mortality rate because of the disease is 2.4% of the population [1]. There are more than 100 types of liver diseases among which cirrhosis is diagnosed when the liver cells are damaged and replaced by non-living scar tissues [2]. One of the most traditional ways to detect liver disease (LD) is to analyze if the liver tissue is abnormal by a specialized radiologist. However, studies show that a decision of accuracy of around 72% can be made by a simple visual interpretation of liver diseases [3]. Since most of the medical centers, hospitals, or diagnosis centers are equipped with modern computer-based machines for testing and diagnosis, early detection of LD is possible for faster cure. Using machine learning algorithms on the lab data, a model can be generated for a much efficient diagnosis. Analysis based on the input and different classification algorithms may give various accuracy rates [4].

According to Gogi and Vijayalakshmi [5], a prognosis of LD was detected using machine learning techniques. For detecting LD LFT dataset was used that has 11 attributes. In the research paper, 5 data mining classification techniques were used and the platform was MATLAB2016. The accuracy found for linear discriminant algorithm was 95.8 % and ROC was 0.93. In the research paper, Midhila *et al.* [6] described a computer-based analysis and classifications for detect 10 types of LD from ultrasound images using some techniques such as segmentation despeckling, feature extraction and gray level difference weights method. The accuracy of classification and segmentation in detecting cysts was 90% and 80% respectively. Kumar and Katyal [7] briefs a method for analyzing LD using data mining techniques. In this research paper, they created a classifications model for diagnosis and to forecast liver problems using 5 data mining algorithms and 1 boosting algorithm. Without boosting, the method's best accuracy of 72.18% was found. Spann *et al.* [8] explained a comprehensive review about LD and transplantation based on machine learning approaches. In the review paper, the authors found that if the patient's data are too large supervised machine learning tools can detect nonalcoholic fatty liver disease (NAFLD) at an early stage. In their research model, Ramkumar *et al.* [9] depict liver cancer prediction based on conditional probability Bayes theorem. In the research WEKA tools and data mining techniques were used to predict liver cancer. It is found that drinking alcohol caused LD based on Bayes theorem. Kefelegn and Kamat [10] presented a survey that used data mining methods to predict and analyze liver problem diseases. Three algorithms such as Naïve Bayes, SVM and C4.5 have been utilized in the study approach. The model has evaluated the performance utilizing a confusion matrix and 10-fold cross-validation for the partitioning of data.

In the proposed study, the symptoms of LD are analyzed and prediction is done to identify if a patient is prone to LD using machine learning models. The classification algorithms used in the process are logistic regression, decision tree (DT), random forest (RF), AdaBoost, k-nearest neighbor (KNN), linear discriminant analysis (LDA), gradient boosting and support vector machine model (SVM). From the research gap of previous studies done on the subject, in this proposed method of prediction, the LASSO feature selection technique is used to identify the features that play a significant role in the occurrence of LD. The accuracy of the proposed model is compared with the existing studies to demonstrate superiority.

2. PROPOSED METHOD

This proposed model takes in the dataset of a number of LD patients from UCI machine learning Repository for the prediction of the disease. Firstly, the raw data is pre-processed to get clean data. From all the attributes in the dataset, the related data are chosen using the LASSO feature extraction method to ensure better accuracy based on only relevant data. Using a 10 fold cross-validation approach and the classification algorithms, data was analyzed. The classification algorithms used in the process to analyze the data are logistic regression, DT, RF, AdaBoost, KNN, LDA, gradient boosting and SVM. The classification results are measured to determine the performance rate. To verify the performance of the system, a performance comparison of the algorithms is done based on accuracy, sensitivity, precision and f1-scores which help to identify the highest performing algorithm as well. A flow of the entire system is illustrated below in Figure 1.

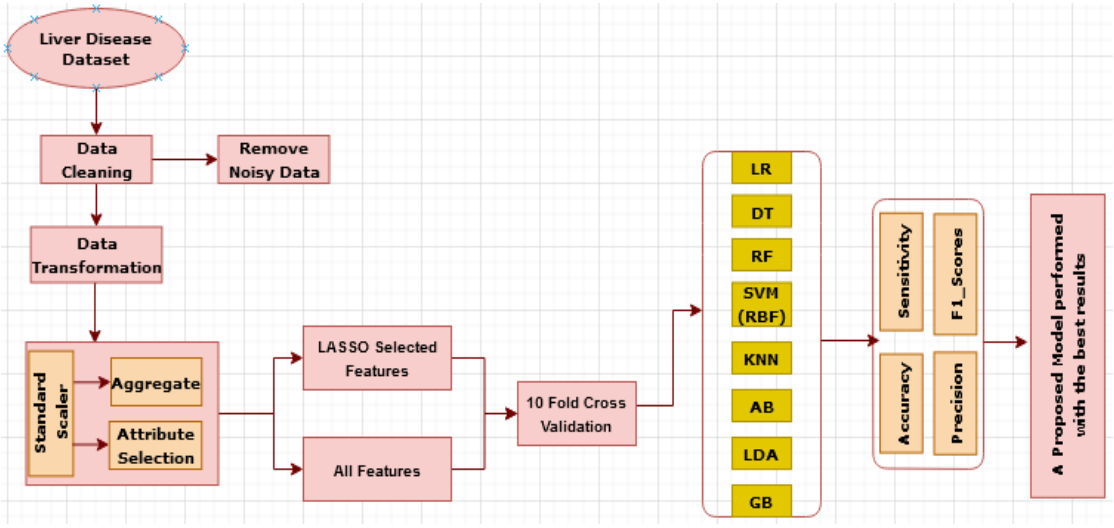


Figure 1. Proposed system diagram

3. RESEARCH METHODOLOGY

3.1. Data collection

For this research, the Indian Liver Patient Dataset (ILPD) is downloaded from UCI machine learning repository [11]. The ILPD dataset has 583 instances and 10 attributes (age of the patients, gender of the patients, total bilirubin, direct bilirubin, alkaline phosphate, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin and albumin and globulin ration) and also a selector field to determine if the subjects are liver patients or not. There are 167 non-LD patients and 416 LD patients that are determined by using the sum of each of the sector fields. Figure 2 shows the data distribution in the dataset. The dataset attribute characteristics are multivariate and attribute characteristics are integer and real [12]. The models are trained and tested on these data and give output for their own which are evaluated for the models' performance.

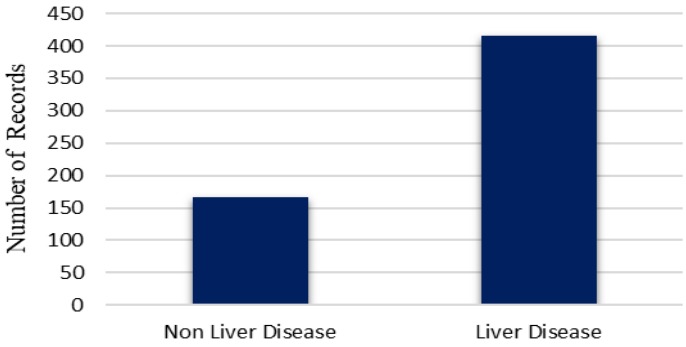


Figure 2. The number of patients in the dataset

3.2. Preprocessing dataset

Data preprocessing is one of the most vital stages in machine learning classification as the cleaner the data, the better the classification result trends to be [13]. The pre-processing techniques applied in the model are described as:

a. *Reduce noisy data:* There are two kinds of data noise in machine learning: attribute noise and class noise. However, for best accuracy in the proposed model, attribute noise is reduced for better accuracy using the panda library.
b. *Data transformation:* Data transformation refers to the process of reorganizing or restructuring raw data. It is used to transform raw data into a suitable format that allows data mining to get strategic information more effectively and quickly.
c. *Standard scalar:* Standard scalar transforms data such that its distribution has a mean value of zero and a standard deviation of one. Aggregate functions perform operations on the column values and return a single value.

3.3. LASSO feature selection

In this proposed method, LASSO techniques are implemented for data fitting and are the best feature selection to reduce overfitting, improves accuracy and reduce training time. LASSO is used to remove unnecessary features from the dataset with high correlation without much loss of information. LASSO techniques minimize the absolute sum of the coefficients. LASSO ridge combines the benefits of regression with subset selection to enhance model understanding and prediction accuracy. If the set of parameters has a strong connection, LASSO selects one of them and reduces the other to zero. This minimizes the variability of the estimate by compressing certain zero coefficients, resulting in a model that is simple to understand [13]. Algorithm 1 shows the working process of LASSO that is implemented in this system.

Algorithm 1: LASSO feature selection

```
Input: Data = {Xi j, Yi j} = 1, 2, … , Ni; Sampling ratio ε ∈ (0, 1);
Output: Relevant features q¯i of the grid
Step 1: number of randomizations T ∈ M; Threshold H ∈ M
Step 2: for k = 1, 2, … , T:
Step 3:   Data = sampling with replacement from data with ratio ε
Step 4:          qi = LASSO-based fingerprint selection using Data
Step 5: end for
```

Step 6: frequency of selection of each feature is calculated according to q_i, k = 1, 2, … , T
Step 7: Return $\bar{q}i$: the set of features selected most frequently

3.4. 10 fold cross-validation

The data set must be divided into a training set and a test set in order to train and test a model. The system uses a 10-fold cross-validation method for this purpose. Algorithm 2 describes the 10 fold cross-validation work procedure.

Algorithm 2: 10 fold cross-validation

Step 1: split dataset randomly into 10 subsets.
Step 2: iterate x = 10 ~> step 3 to step 5
Step 3: for x^{th} iteration, consider x^{th} subset as test set and rest as Training sets
Step 4: train the model on the training sets
Step 5: test the model on x^{th} test set
Step 6: take the Mean of the ten results as the final output

3.5. Data classification

a. *Decision tree:* The decision tree is the most popular supervised learning algorithm for prediction. As the name suggests, the algorithm is formed in a tree structure with the root node, branches and leaf nodes that indicate attributes, conditions and outcomes respectively [14]. Entropy as denoted in (1) shows the homogeneity as well as the purity of a dataset, and information gain is the change in an input's entropy, which is usually a reduction [15].

$$\mathrm{E\ (D)\ =\ -P\ (positive)\ log2\ P\ (positive)\ -\ P\ (negative)\ log2\ P\ (negative)} \quad (1)$$

b. *Random forest:* RF is a constituent of multiple decision tree algorithms. It can be used for both classification and regression. Random forest prevents the model from overfitting to give better predictions [16]. According to the data provided, this system ranges from the lower limit of b=1 to the higher limit of B. The unknown samples x' are generated by averaging the predictions $\sum_{b=1}^{B} fb(x')$ from each tree on x' as stated in (2),

$$j = \frac{1}{B}\sum_{b=1}^{B} fb(x') \quad (2)$$

c. *K-Nearest Neighbor:* KNN algorithm works based on similar things that exist in close proximity. Some advantages of the KNN algorithm for instance: simple and easy to implement [17]. The KNN algorithm uses Euclidean distance in (3) where p and q are two different data points.

$$\mathrm{d(p,q)} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (3)$$

d. *Support vector machine:* The data points are split into two classes by a hyperplane affected by the support vectors using the SVM supervised classification method with a RBF kernel [18]. The Euclidean distance is used to calculate the distance between the support vectors and the hyperplane, as shown in (4).

$$\beta 0 + \beta 1X1 + \beta 2\,X2 + \ldots + \beta nXn = 0 \quad (4)$$

Where β0, β1, β2 … βn represent hypothetical values and Xn represent data points in the n-dimensional sample space. The original goal of creating SVM was to handle a two-class classification issue; however, it was subsequently adjusted for multi-class situations.

e. *AdaBoost classifier:* It is an ensemble model. Constructed using n numbers of decision tree [15]. The incorrectly classified data after training in the first decision tree is passed to the next tree for classification until n^{th} tree to get the most accurate prediction as shown in (5).

$$\mathrm{Weight\ (xi)\ =\ 1/n} \quad (5)$$

The frequency of training instances is represented by n, and the i^{th} training instance is represented by xi. The decision stump generates an output for each input variable.

f. *Logistic regression:* This algorithm is a classification technique that is important for machine learning and data mining applications. It is commonly used because of two-class classification and it has two types of regression such as Binary and Multi-linear function fails the class. It analysis result based on the

concept of probability. Input dataset, D={(x_1, y_1), (x_2, y_2),……..(x_p, y_p), Base learner L and Number of base learners, B.

Process:

For b = 1 to B:

hb = L (Db) // Train a base learner hb

end

Output: $H(x) = \left(\sum_{b=1}^{B} h_b(x)\right) / B$

g. *Linear discriminant analysis:* As the name implies, the model reduces dimension in the dataset yet keeps enough information for classification [19]. It uses the information of the kept dimensions and constructs a next axis to minimize the variance and distance between the classes as illustrated in Figure 3.

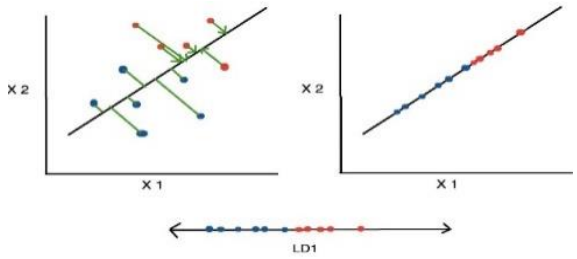


Figure 3. Working process of LDA

h. *Gradient boosting:* To learn gradient boosting first we need to know about boosting, boosting is a method that converts week learners to strong learners. Here each tree is a fit on a modified version of the original data set. The boosting technique differs from traditional machine learning in that function space does not allow for optimization. After m^{th} iterations, the optimal function F(X) is found, which is computed using (6):

$$F(X) = \sum_{i=0}^{m} f_i(x) \quad (6)$$

Where fi (x) (i=1, 2…., M) indicates feature increments, the fi (x) = − ρi x gm(X).

3.6. Performance measurements

In this research paper, we use confusion metrics because of its best and easiest way to calculate the performance of a classification result that has two or more types of classes for output [20]. Using the matrices (TP, TN, FP, FN), the performance of the models is measured using (7) to (10). Table 1 and 2 shows the construct of the confusion matrix.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (7)$$

$$\text{Sensitivity or Recall} = \frac{\text{TP}}{\text{TP+ FN}} \quad (8)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (9)$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision+recall}} \quad (10)$$

4. RESULTS AND DISCUSSION

In the system, the LASSO feature selection technique was used to determine the important features for the classification. Figure 4 illustrates and rates the importance of each feature in the dataset. It is seen that only six features are needed for classification according to LASSO feature selection which aids in achieving a higher accuracy rate.

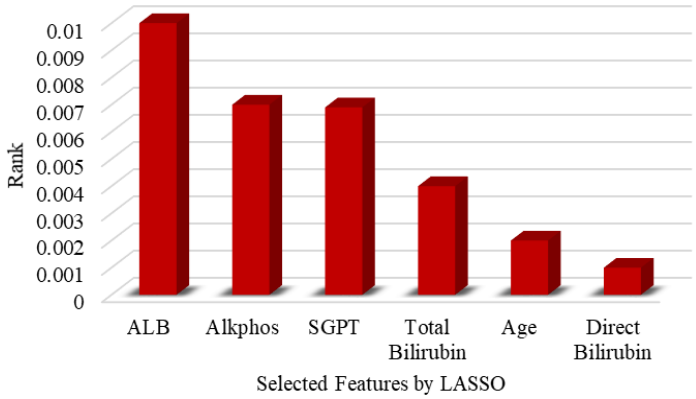


Figure 4. Feature importance using lasso technique

To verify that using LASSO feature selection, the accuracy of the models has increased drastically, a comparison between the accuracy of classification of the algorithms with and without using the feature selection technique is done. In Figure 5, the accuracy rate of without and with LASSO rate is depicted. It is observed that when used all features, the logistic regression model shows the highest accuracy of 77.1428% whereas, with LASSO, the decision tree model shows an accuracy of 94.285%.

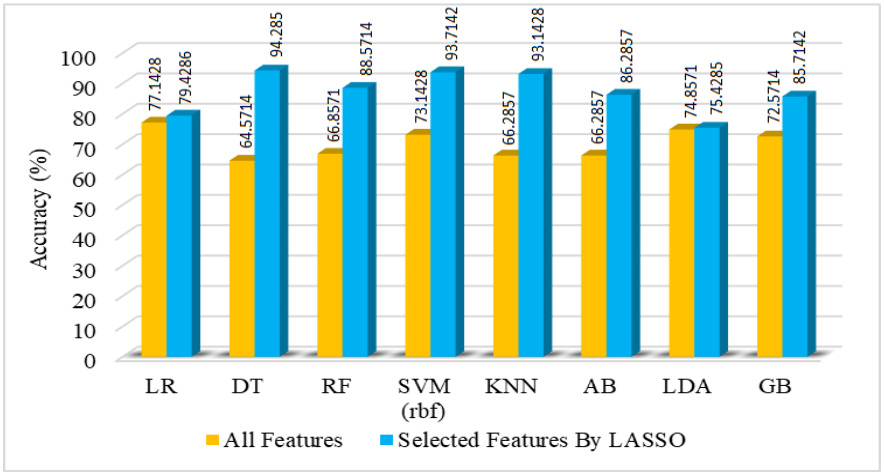


Figure 5. Accuracies of all features and selected features by the LASSO approach

The performance of the models is assessed using the confusion matrix for all of the characteristics in the dataset. The linear regression method has already been shown to be better in terms of accuracy for all characteristics. Precision scores, recall scores, and f1-scores are computed and shown in Table 1 among the models using the performance measure matrices. It is observed that the linear regression has the highest precision, recall and f1-score as well with 76.7%, 75.3% and 75.1% respectively.

Table 1. Performance measures of all models on the full features of LD

Algorithms	Precision	Recall	F1-Score
LR	76.7%	75.3%	75.1%
DT	63.3%	63%	63.4%
RF	67.4%	67%	66.8%
SVM(rbf)	72.9%	71%	71%
KNN	66.4%	65.6%	65%
AB	66.4%	65.6%	65.7%
LDA	73.4%	73%	73.3%
GB	71.2%	70.9%	71%

Using a confusion matrix, the performance of the models is measured. It is already established that the decision tree algorithm shows superior accuracy after feature extraction. Among the models, with the performance measure matrices, precision scores, recall score and f1-score are calculated and illustrated in Table 2. It is observed that the decision algorithm has the highest precision, recall and f1-score as well with 92%, 99% and 95.3% respectively.

Table 2. Performance measures of the introduced models on LASSO features

Algorithms	Precision	Recall	F1-Score
LR	77%	95%	85%
DT	92%	99%	95.3%
RF	85%	98%	91.2%
SVM(rbf)	92%	98%	95%
KNN	90%	99%	95%
AB	88%	90%	88.7%
LDA	74%	92%	82.3%
GB	84%	95%	89.8%

Table 3 illustrates the comparison among various studies done on the topic in recent years with the proposed model of the paper. It can the observed from the comparison that the proposed model shows much higher accuracy compared to the studies done using other machine learning models to predict LD patients.

Table 3. Comparison result with existing system

Author	Year	Dataset	Models	Accuracy
Ghosh *et al.* [1]	2020	ILPD	DT	94.29 %
Karasu *et al.* [21]	2018	ILPD	LG	73.97 %
Singh *et al.* [22]	2019	ILPD	LG	72.50 %
Thaiparnit *et al.* [23]	2018	Liver Disorder	RF	75.76 %
Rahman *et al.* [24]	2019	ILPD	LG	75%
Kumar and Thakur [25]	2020	BUPA, ILPD	Fuzzy-NWKNN	78.46%
Rabbi *et al.* [26]	2020	ILPD	AdaBoost	92.19%
Poonguzharselvi *et al*. [27]	2021	UCI repository	Random Forest	84%

5. CONCLUSION

The system proposed in the paper contributes to the field of medical science by helping to identify LD in a patient from certain data at an early stage that will allow to start the treatment and cure the disease before it becomes fatal. In order to do so data of the age of the patients, total bilirubin, direct bilirubin, alkaline phosphate, albumin are needed. This is determined by the LASSO method. Using classification algorithms, it is observed that the decision tree algorithm has the most accurate prediction amongst seven other classifiers i.e. RF, LR, SVM, KNN, LDA, AdaBoost and gradient boosting. It has an accuracy rate of 94.285% followed by SVM with an accuracy rate of 93.7142% The decision tree model also has precision, sensitivity and f1-score of 0.92, 1.00 and 0.96 respectively. Using this proposed model in the future, other diseases such as cancer, parkinson, alzheimer can be predicted as well.

REFERENCES

[1] P. Ghosh, A. Karim, S. T. Atik, S. Afrin and M. Saifuzzaman, "Expert model of cancer disease using supervised algorithms with a LASSO feature selection approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, 2020, doi: 10.11591/ijece.v11i3.pp2631-2639.

[2] P. Ghosh, S. Azam, A. Karim, M. Jonkman and MD. Z. Hasan, "Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases," *5th International Conference on Information System and Data Mining (ICISDM2021)*, Silicon Valley, USA, pp. 14–20, May 2021, doi: 10.1145/3471287.3471297.

[3] M. Abdar, M. Zomorodi-Moghadam, R. Das and I-Hsien Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Systems with Applications*, vol. 67, pp/ 239–251, January 2017, doi: 10.1016/j.eswa.2016.08.065.

[4] P. Kaur, R. Kumar and M. Kumar, "A healthcare monitoring system using random forest and internet of things (iot)," *Multimedia Tools and Applications*, pp. 19905–19916, 2019, doi: 10.1007/s11042-019-7327-8.

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

[6] M. Midhila, K. R. Krishnan and R. Sudhakar, "A study of the phases of classification of liver diseases from ultrasound images and gray level difference weights based segmentation," *2017 International Conference on Communication and Signal Processing (ICCSP)*, 2017, pp. 1582-1587, doi: 10.1109/ICCSP.2017.8286655.
[7] S. Kumar and S. Katyal, "Effective Analysis and Diagnosis of Liver Disorder by Data Mining," *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018, pp. 1047-1051, doi: 10.1109/ICIRCA.2018.8596817.
[8] A. Spann *et al.*, "Applying Machine Learning in Liver Disease & Transplantation: A Comprehensive Review. Hepatology," *Hepatology,* 2020, doi: 10.1002/hep.31103.
[9] N. Ramkumar, S. Prakash, S. A. Kumar and K. Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem," *2017 International Conference on Computer Communication and Informatics (ICCCI)*, 2017, pp. 1-5, doi: 10.1109/ICCCI.2017.8117752.
[10] S. Kefelegn and P. Kamat, "Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey," *International Journal of Pure and Applied Mathematics*, vol. 118, pp. 765-770, 2018.
[11] UCI Machine Learning Repository, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/00225/, (Last Accessed 28-4-2021).
[12] D. Ramesh and Y. S. Katheria, "Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach," *Health and Technology*, vol. 9, no. 4, pp. 533-545, 2019, doi: 10.1007/s12553-019-00299-3.
[13] P. Ghosh *et al*., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in *IEEE Access*, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
[14] Z. Tasnim *et al.*, "Deep Learning Predictive Model for Colon Cancer Patient using CNN-based Classification*," International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 8, 2021, doi: 10.14569/IJACSA.2021.0120880.
[15] S. M. Basha, D. S. Rajput and V Vandhan, "Impact of gradient ascent and boosting algorithm in classification," *International Journal of Intelligent Engineering and Systems (IJIES),* vol. 11, no. 1, pp. 41–49, 2018, doi: 10.22266/ijies2018.0228.05.
[16] P. Ghosh, S. Azam, K. M. Hasib, A. Karim, M. Jonkman and A. Anwar, "A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast Cancer," *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9534293.
[17] C. Yu, H. Chen, Y. Li, Y. Peng, J. Li and F. Yang, "Breast cancer classification in pathological images based on hybrid features," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 21325–21345, 2019, doi: 10.1007/s11042-019-7468-9.
[18] 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.
[19] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy and M. Jonkman,"A Comparative Study of Different Machine Learning Tools in Detecting Diabetes," *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, vol. 192, pp. 467-477, 2021, doi: 10.1016/j.procs.2021.08.048.
[20] F. M. J. Mehedi Shamrat, S. Chakraborty, M. M. Billah, M. A. Jubair, M. S. Islam and R. Ranjan, "Face Mask Detection using Convolutional Neural Network (CNN) to reduce the spread of Covid-19," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1231-1237, doi: 10.1109/ICOEI51242.2021.9452836.
[21] K. Thirunavukkarasu, A. S. Singh, M. Irfan and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms," *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1-3, doi: 10.1109/CCAA.2018.8777655.
[22] J. Singh, S. Bagga and R. Kaur, "Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques," *International Conference on Computational Intelligence and Data Science (ICCIDS)*, vol. 167, pp. 1970–1980, 2019, doi:10.1016/j.procs.2020.03.226.
[23] S. Thaiparnit, N. Chumuang and M. Ketcham, "A Comparitive Study of Clasification Liver Dysfunction with Machine Learning," *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1-4, doi: 10.1109/iSAI-NLP.2018.8692808.
[24] F. M. J. M. Shamrat *et al.,*"Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm*," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 23, no. 1, pp. 463-470, 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.
[25] P. Kuma and R. S. Thakur, "Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach," *Multimedia Tools and Applications*, vol. 80, pp. 16515–16535 2020, doi: 10.1007/s11042-019-07978-3.
[26] M. F. Rabbi, S. M. Mahedy Hasan, A. I. Champa, M. AsifZaman and M. K. Hasan, "Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study," *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, 2020, pp. 111-116, doi: 10.1109/ICAICT51780.2020.9333528.
[27] B. Poonguzharselvi, M. M. A. Ashraf, V. V. S. S. Subhash and S. Karunakaran, "prediction of Liver Disease Using Machine Learning Algorithm and Genetic Algorithm," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4, 2021.