A HYBRID APPROACH FOR THE PREDICTION OF LIVER DISORDER USING DATA MINING

Himanshi Bansal[1], Kapil Sharma[2] and Goldendeep Kaur[3]

[1]Research Scholar, GNDEC, Ludhiana, Punjab

[2,3]Assistant Professor, GNDEC, Ludhiana, Punjab

ABSTRACT

With the increase in the number of patients with liver disorder it has become a fatal disease in many countries. These disorders of liver need clinical care by professionals in healthcare. Data mining provides with essential methodology for computing applications in the domain of medicine. Medical data mining uses a set of approaches that extract valuable patterns from the human services databases to help doctors pick the best diagnosis. Many classifiers had been used to predict the liver disorder. Hybrid algorithms for data mining are a logical combination of multiple pre-existing techniques to enhance performance and provide better results. In the proposed work, a hybrid algorithm is introduced which uses the concept of clustering and classification for prediction analysis. We have analyzed with real patient dataset for constructing hybrid approach to predict liver disorder taken from the UCI repository. The algorithms used in this work are K-means clustering, Random Forest classifier. This hybrid algorithm is implemented using Jupyter Notebook. This approach is compared with existing classifiers like KNN, LR etc using the confusion matrix parameters such as precision, recall, f1-score and accuracy. The hybrid approach shows better results than other algorithms for prediction analysis of liver disorder.

Keywords: Data Mining, Liver disorder, Clustering, Classification, Random forest, K-means

1. INTRODUCTION

Data mining often is characterized as the method of extracting useful data and patterns from a broad range of raw dataset. This includes identifying data trends or patterns utilizing one or more tools in huge quantities of results. It has uses in many areas such as research, healthcare and analysis. Since data mining is the method of exploration and detection of patterns where large amount of data is involved, some efficient early prediction systems related to healthcare have appeared from the medical datasets [1].

In medical field, enhancing the standard of healthcare is best characterized by the guiding factors that influence it which includes data on healthcare meaning the dataset can be considered as the basis of the system in the plan of process improvement for every patient. Utilizing the data mining techniques to extract knowledge from the medical records or datasets will contribute to the discovery of disease occurrence, development, identification and give useful details to determine the sources of diagnostic procedures depending on the factors existing within healthcare. Data mining approach could also be used in the exploration of the information cycle to classify diseases. Accordingly, it will revel invisible connections, identifying patterns between data which results to better and improved diagnostic identification.

This paper is a study of liver disease explained further. Liver is one of the main organs of the body weighing approximately 3 lbs. on the right side of the abdominal cavity beneath the diaphragm, liver is located. The portal veins and hepatic artery are the two veins responsible for transporting blood to liver. These provides oxygen rich blood to heart vessels. They branch out continuously within liver which ends in utterly small capillaries. Any liver disease can result in liver failures, inflammation, weak functioning and may affect alternate organs of the body. Due to its functions, liver forms an important part of the body and the disease needs to be detected at an early stage. Applying data mining techniques, can help to build this prediction system for liver disorder.

The objective of this research work is to propose a hybrid approach using techniques in data mining to develop a system which helps to identify whether or not the patient has liver disorder. Model of predictive analysis and various algorithms are used for making predictions for this purpose. The process of this model is held in four stages. In initial stage, pre-process the raw data. In second stage, convert the processed data into an adaptable form that can be applied to the model. Train the model in third stage using the dataset. Forth stage makes predictions using the learning model and review them as required.

2. LITERATURE SURVEY

[2] in this article, researchers suggested a model for demanding liver problems to isolate the beneficial data from the patient using Artificial Neural Network technique. The suggested approach used derived functionality for description of the function using M-PSO and Artificial Neural Network algorithm. SPARK tool was used for implementation as it supports big data mining and artificial intelligence. They reviewed the display of the proposed research and compared with current techniques which incorporated C5.0 with CHAID. The paper reflected on description of the chosen features. The test findings indicated that the new approach suggests a skillful management of liver disease.

[3] in this paper, researchers used various types of learning which are categorized in machine learning on the liver disease dataset. The types of learnings discussed in this paper were supervised learning, unsupervised learning and reinforcement learning. Supervised learning involves mapping function which is learned by an algorithm and helps to map the input variable to the output variable. Unsupervised learning understands the structure of datasets and has no corresponding output variable. Reinforcement learning allows the system or machine to determine the unique behavior’s within a context to maximize its performance. This paper focused on machine calculations. After examining various related concepts, the algorithms KNN and SVM can give an improved system for liver disorder dataset.

[4] in this paper, researchers explored various techniques of decision trees for predicting early liver disease. The techniques used were J48, Random Forest, LMT, Random Tree, Decision Stump, Rep tree and Hoeffding Tree. The dataset of liver patients was evaluated by using these techniques and their performance was compared by calculating seven performance metrics accuracy, mean absolute error, precision, recall, F-measure, Kappa Statistics and runtime. The objective of this paper was to find the best decision tree algorithm for the prediction purpose. The comparison analysis performed on decision tree algorithms showed that Decision Stump receives the highest accuracy of 70.67% and the execution time is least for Random Tree and Decision Stump. From the analysis result, it was concluded that Decision Stump performs better as compared to other techniques.

[5] in this article, researchers aim to include a survey study of liver disease prediction using various approaches of machine learning. Machine learning algorithms are especially helpful in providing patient condition critical quantities, constant knowledge, detailed analysis, clinical details and much more to physicians. The key objective was to show the value of various predictive disease classification algorithms especially related to liver disorder such as NN, Decision Tree, KNN, Random Forest, SVM, Logistic Regression. The datasets used are associated to liver disorders such as hepatitis and carcinoma.

[6] in this paper, authors demonstrated how to identify the dataset of patients with liver disease to predict accurate liver diagnosis utilizing different data mining algorithms. To access the output of such classifiers, five classification techniques were used KNN, C5.0, K-means, Naïve Bayes, Random Forest and C5.0 with adaptive boosting. The implementation part for this work was done in R-studio software using the R language. The highest accuracy was achieved by Random Forest algorithm of 72% without making any changes to the algorithm. However, C5.0 increases the accuracy to 75% when adaptive boosting was applied. K-means obtained the highest precision and KNN obtained highest recall value.

[7] in this research article, authors intend on providing a survey on liver disease prediction analysis using different data mining classifiers. They experimented by using various classification algorithms in data mining to generate efficient results compared to previous studies related to liver disorder prediction. The experiment was performed in weka tool on a dataset with 583 instances and 10 attributes. The proposed work used three algorithms SVM, Naïve Bayes and C4.5 decision tree. The method of K-fold cross validation is used for partitioning the data. They compared the algorithms on the basis of confusion matrix. They propose that we may use a hybrid approach to generate better results and models which perform more accurately.

[8] in this research article, authors proposed that there exists a number of liver diseases which requires medical attention of the practitioners. They performed a comparison between two algorithms Naïve Bayes and FT Tree to see which performs better. The dataset used has 54 instances and 12 attributes. The data was pre-processed in Weka tool where comparison results were obtained by implementing the given algorithms. Among the two classification algorithms, Naïve Bayes gives an accuracy performance of 75.54% which is better than FP growth algorithm. Accuracy is calculated with the use of confusion matrix.

3. DESCRIPTION OF ALGORITHMS USED

3.1 KNN

K-Nearest Neighbor classifier is a supervised learning algorithm in which a specific collection of data is split into many clusters as defined by the user [9]. Its implementation is easy and is used for both regression and classification related problems. This algorithm believes that the items that are similar, happen to be nearby i.e. similar things are close to one another. It is sometimes considered as the lazy learner algorithm since it doesn’t learn directly from training dataset rather it holds the dataset and conducts an operation on the data at the period of classification. New data points are classified based on the measurements of similarity like distance function. By a majority vote of its neighbors, classification takes place. The point is allocated to the class with its nearest neighbors. Changing the value of k, changes the algorithm’s accuracy.

3.2 SVM

Support vector machine is a specialized supervised learning classifier found in statistical learning used for both linear and non-linear dataset classification. With the help of non-linear mapping function, it converts the initial dataset into meaningful information. In this dimension it is searching for a linear hyperplane that separates the data points. A hyperplane is an optimal boundary for judgement. SVM constructs these hyperplanes using support vectors. Hyperplane separates the data into classes using a suitable non-linear mapping function [10]. SVM is a reliable and precise classification algorithm in which analysis is focused on sequential optimization programming and is expensive, as to solve these quadratic problems one needs mathematical functions and also include complex calculations which are time intensive [11].

3.3 Random Forest

Random forest is an algorithm of supervised learning. It creates a forest which is an ensemble of decision trees, taught by the process of bagging approach. The basic theory for the approach of bagging is that a mixture of learning models would maximize the total output. Since it is an ensemble, it brings more randomness in the model as the tree grows. Rather than searching for the most appropriate feature, it looks for the most contributing feature when a node split. Choosing the best feature out of all the random set of features, it contributes in generating an efficient model. This algorithm constructs decision trees on dataset, receives the output with majority votes. This approach is more efficient than only one decision tree and eliminates overfit by combining the output.

3.4 K-means Clustering

K-means clustering, an unsupervised learning algorithm is the simplest and easiest clustering algorithm to tackle the problem solving. This approach applies a straightforward and clear method of defining a given dataset in a set of selected clusters (value of k gives the number of clusters). The aim is to define a centre for each cluster. These k centres are placed at specific places and far away from each other so they produce different results. Each data point is allocated to the nearest centre using the distance function like Euclidean distance. this algorithm minimizes the distance between the data points of a cluster from their centroid. The algorithm keeps on renewing these centroids until no more adjustments can be made i.e. the distance has been minimized between the points and their centres. Choosing the k value is most important in this approach.

4. HYBRID APPROACH

Hybrid usually means a combination of two or more elements which may be similar or are different in their properties. Different elements have different features but if they are combined, the new element may have both the features. In case of data mining, hybrid approach means a combination of two or more algorithms having their own advantages. The algorithms when combined together generates new results which may be more efficient and accurate than using those algorithms individually. In the proposed work, a combination of clustering and classification is used for predictive analysis of liver disorder. For clustering, k-means clustering algorithm is used and for classification, random forest classifier is used. K-Mean clustering acts as a pre-classification process grouping items into number of disjoint clusters depending on the features. It is used for feature extraction which generates new k-means features. Random forest algorithm is used for classification purpose. The dataset is trained using this classifier. This hybrid approach generates results with more accuracy when compared with the existing algorithms on the basis of various parameters.

5. METHODOLOGY

The proposed work is implemented using python programming language in Jupyter Notebook. All the steps of implementation for the methodology are applied here. The dataset used for training the system is taken from the UCI repository of machine learning [12]. The Figure 1 shows the flowchart which contains the steps followed for the methodology of proposed work.

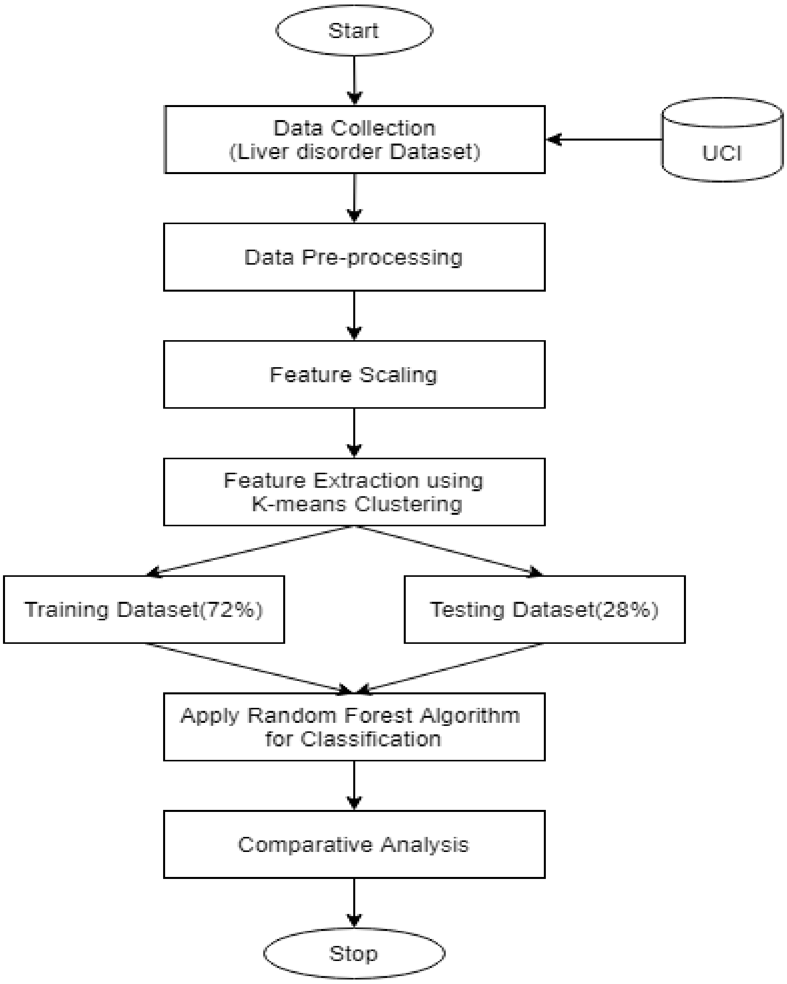

Start
Data Collection (Liver disorder Dataset)
UCI
Data Pre-processing
Feature Scaling
Feature Extraction using K-means Clustering
Training Dataset(72%)
Testing Dataset(28%)
Apply Random Forest Algorithm for Classification
Comparative Analysis
Stop

Figure 1 Flowchart of Methodology

5.1 Data Description

For this study, Indian Liver Patient Dataset is used downloaded from the UCI repository of machine learning for building the model. This dataset has 583 records of patients and 11 attributes. Out of this, 416 instances are of patients with liver disorder and 167 instances are of non- liver disorder. Description of the attributes is given in Table 1.

Table 1: Attribute Description

Sr no.	Attribute Name	Data Type
1.	Age	Number
2.	Gender	Category
3.	Total_Bilirubin	Number
4.	Direct_Bilirubin	Number
5.	Alkaline_Phosphotase	Number
6.	Alamine_Aminotransferase	Number
7.	Aspartate_Aminotansferase	Number

8.	Total_Proteins	Number
9.	Albumin	Number
10.	Albumin_and_Globulin_Rati	Number
11.	Dataset	Number

5.2 Data Pre-processing

Data pre-processing is an important step that makes the data ready to be used for training the model. The data is cleaned and pre-processed so that it can used for implementation. In this, data is analyzed to see if there are any missing values. The missing values are filled by the mean value of that column attribute. Also, the attribute with categorical values are converted into numerical values like gender for easy analysis.

5.3 Feature Scaling

Feature scaling is important as it helps to normalize the set of variables or values in the data. The data contains raw values which vary widely and may not allow some algorithms to function properly. So, for this study standard scaler is used that brings the values of all features in a particular range.

5.4 Feature Extraction

In this proposed work, k-means clustering is used for feature extraction. This clustering approach generates new features and add it to the dataset. Using these k-means features makes the performance of the model better.

5.5 Training and Testing data

The pre-processed dataset is divided into training (72%) and testing (28%) dataset. The training dataset is used to train the model and algorithms. The testing dataset is used to test the model about accurately and efficiently the results are generated.

5.6 Comparative Analysis

Comparative analysis is done to compare the performance of the proposed approach with the existing techniques. It proves that the proposed work is more accurate, efficient and can be used for predictive analysis. In this work, confusion matrix parameters are used for comparison. Confusion matrix is a table frequently used to interpret a model's performance applied on a dataset with known true values. A confusion matrix provides an analysis of the results obtained of a prediction over a classification issue. This matrix generates four values- TP (True Positive), FN (False Negative), TN (True Negative), FP (False Positive). Various parameters are derived from confusion matrix which are used for comparison. The parameters used in this study are as follows and are calculated for each technique:

Accuracy – It measures the percentage to which the measured values are conforms to the actual true values.

$$\text{Accuracy} = (TP+TN) / (TP+FP+TN+FN) \tag{1}$$

Precision – It calculates the percentage of relevant results obtained.

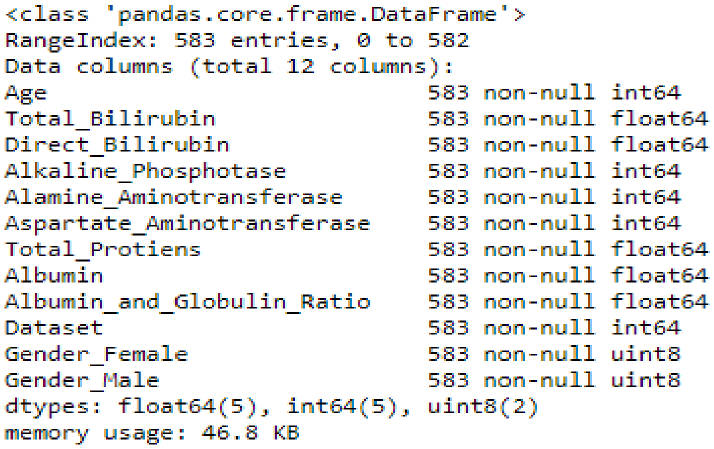
$$\text{Precision} = TP / (TP+FP) \tag{2}$$

Recall – It calculates the percentage of total relevant values which are classified by the algorithm correctly.

$$\text{Recall} = TP / (TP+FN) \tag{3}$$

6. RESULTS AND DISCUSSIONS

The operations and the methodology steps mentioned in previous section are performed using python language in Jupyter Notebook. Jupyter Notebook is a client server application platform. The results obtained in the proposed work are mentioned in this section.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 12 columns):
Age                             583 non-null int64
Total_Bilirubin                 583 non-null float64
Direct_Bilirubin                583 non-null float64
Alkaline_Phosphotase            583 non-null int64
Alamine_Aminotransferase        583 non-null int64
Aspartate_Aminotransferase      583 non-null int64
Total_Protiens                  583 non-null float64
Albumin                         583 non-null float64
Albumin_and_Globulin_Ratio      583 non-null float64
Dataset                         583 non-null int64
Gender_Female                   583 non-null uint8
Gender_Male                     583 non-null uint8
dtypes: float64(5), int64(5), uint8(2)
memory usage: 46.8 KB
```

Figure 2 Pre-processed dataset

The Figure 2 shows that the dataset is cleaned and analyzed. Data pre-processing has been performed where missing values have been filled by the mean value of the attribute, categorical data values like gender is converted into numerical data values, feature scaling has been performed to normalize or standardize the dataset values.

```
Number of patients diagnosed with liver disease:  416
Number of patients not diagnosed with liver disease:  167
```

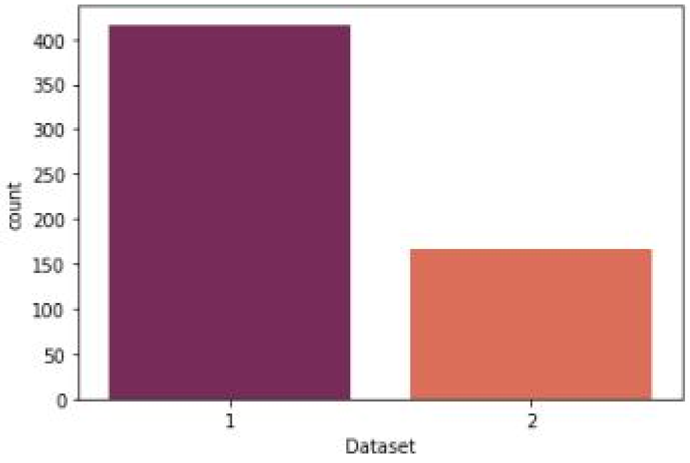


Figure 3 Data Visualization

Figure 3 visualizes the dataset used for prediction of liver disorder. The dataset contains 416 records of patients with liver disorder and 167 records of patients without liver disorder.

Table 2: Results obtained

Parameters / Algorithms	Accuracy	Precision	Recall	F1-score
KNN	72.05	80.98	83.78	82.35
LR	73.23	78.57	88	83.01
SVM	71.12	71.14	98.98	82.78
Random Forest	92.06	95.18	94.21	94.70
Hybrid Approach	95	96.56	95.57	96.06

Table 2 shows the results of performance parameters such as accuracy, precision, recall and f1-score obtained for proposed hybrid approach and existing algorithms after implementation.

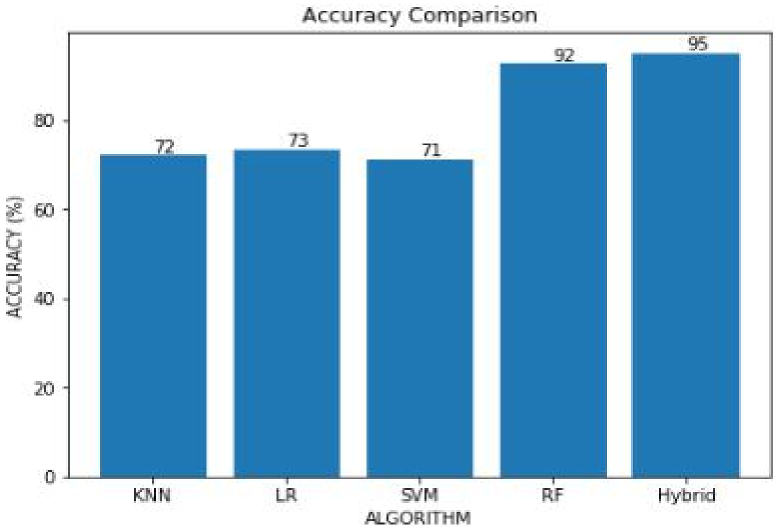


Figure 4 Comparison of algorithms in terms of Accuracy

The Figure 4 shows the comparison of existing algorithms and proposed approach in terms of accuracy performance parameter. The accuracy value obtained of proposed hybrid approach is highest among other algorithms.

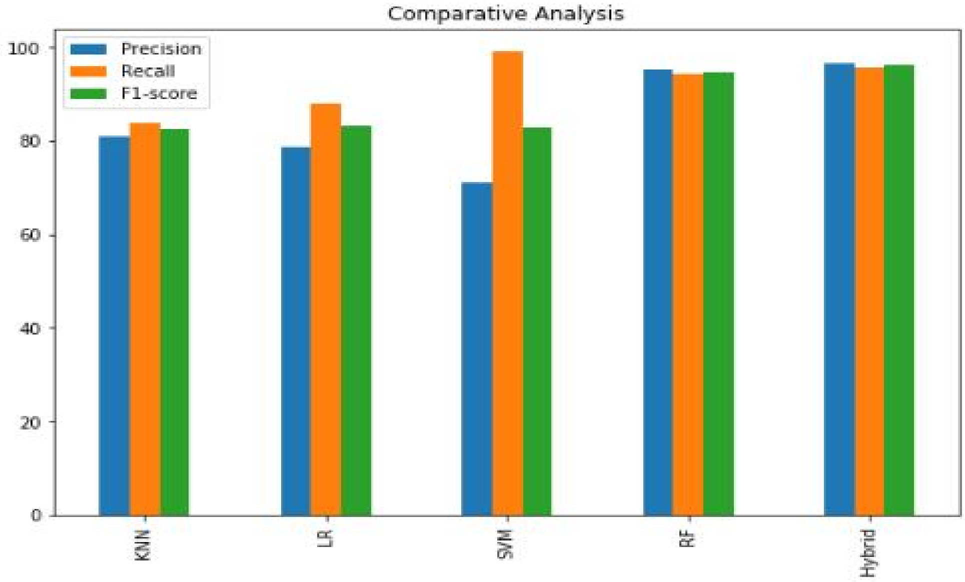


Figure 5 Comparison of hybrid approach with other algorithms

The graph in Figure 5 shows that the values of parameters for hybrid approach are better than other algorithms. The performance analysis of the proposed work is better than the existing algorithms for predicting liver disorder.

7. Conclusion

Data mining is the method by which one can study the raw data easily and draw out important patterns which can be useful for further implementations. It provides with various classification techniques which are useful for predicting liver disorder. In this study, we proposed a hybrid approach using clustering and classification to develop a model for prediction liver disorder and is more accurate and generates better results. This approach uses k-means clustering to obtain attributes from a large dataset and then applies the random forest for classification to generate a model for predictive analysis. The proposed approach shows better results when compared with existing algorithms which are KNN, LR etc. The study shows that hybrid approach obtained an accuracy of 95% on the same dataset for prediction of liver disorder which is higher than other algorithms.

References

[1] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection," Informatics Med. Unlocked, vol. 17, no. October, p. 100255, 2019, doi: 10.1016/j.imu.2019.100255.

[2] L. Anand and V. Neelanarayanan, "Liver disease classification using deep learning algorithm," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 5105–5111, 2019, doi: 10.35940/ijitee.L2747.1081219.

[3] H. Mehtaj Banu, "Liver disease prediction using machine-learning algorithms," Int. J. Eng. Adv. Technol., vol. 8, no. 6, pp. 2532–2534, 2019, doi: 10.35940/ijeat.F8365.088619.

[4] N. Nahar and F. Ara, "Liver Disease Prediction by Using Different Decision Tree Techniques," Int. J. Data Min. Knowl. Manag. Process, vol. 8, no. 2, pp. 01–09, 2018, doi: 10.5121/ijdkp.2018.8201.

[5] A. N. Hanoon, A. A. Abdulhameed, S. R. Al Zaidee, and Q. S. Banyhussan, "Machine learning techniques on liver disease - A survey," Int. J. Eng. Technol., vol. 7, no. 4.20 Special Issue 20, pp. 485–490, 2018, doi: 10.14419/ijet.v7i3.12.16165.

[6] S. Kumar and S. Katyal, "Effective Analysis and Diagnosis of Liver Disorder by Data Mining," Proc. Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2018, pp. 1047–1051, 2018, doi: 10.1109/ICIRCA.2018.8596817.

[7] S. Kefelegn, "Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey," vol. 118, no. 9, pp. 765–770, 2017.

[8] S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classfication," 4th Natl. Conf. Adv. Comput. Appl. Technol., no. May, pp. 1–3, 2014.

[9] T. R. Baitharu and S. K. Pani, "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset," Procedia Comput. Sci., vol. 85, no. December, pp. 862–870, 2016, doi: 10.1016/j.procs.2016.05.276.

[10] M. S. D. Dr. S. Vijayarani1, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," Int. J. Sci. Eng. Technol. Res., vol. 4, no. 4, pp. 816–820, 2015.

[11] M. Banu Priya, P. Laura Juliet, and P. R. Tamilselvi, "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms," Int. Res. J. Eng. Technol., vol. 5, no. 1, pp. 206–211, 2018.

[12] "UCI Official site." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ ILPD+(Indian+Liver+Patient+Dataset).

[13] P. M. Barnaghi, V. A. Sahzabi, and A. A. Bakar, "A Comparative Study for Various Methods of Classification," Int. Conf. Inf. Comput. Networks, vol. 27, no. Icicn, pp. 62–66, 2012.

[14] R. S. Ishi, K. Road, and N. T. Naka, "A Review on algorithms of Liver Disease Diagnosis 1," vol. 1, no. 5, pp. 811–816, 2016.

[15] M. S and M. E, "An Analysis on Clustering Algorithms in Data Mining," Int. J. Comput. Sci. Mob. Comput., vol. 3, no. 1, pp. 334–340, 2014.

[16] A. S. Romana, "A Comparative Study of Different Machine Learning Algorithms for Disease Prediction," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 7, no. 7, p. 172, 2017, doi: 10.23956/ijarcsse/v7i7/0177.

[17] P. Rajeswari and G. S. Reena, "Analysis of Liver Disorder Using Data mining Algorithm," Glob. J. Comput. Sci. Technol., vol. 10, no. 14, pp. 48–52, 2010.

AUTHOR

Himanshi Bansal received Bachelor of Engineering in Computer science and engineering degree from Chandigarh University, Mohali (Punjab) in 2018. She is currently pursuing Master of Technology in Computer Science & engineering from Guru Nanak Dev Engineering College, Ludhiana (Punjab). Her main domain for research is data mining and focuses on development of a hybrid approach for predicting liver disorder using clustering and classification algorithms used in data mining.