Liver Diseases Prediction

Abstract—Bangladesh is one of the most densely populated countries in the world and because of that it faces significant challenges in managing the health of its population. With limited healthcare resources and a high burden of infectious and non-communicable diseases, liver disease has emerged as a major public health concern. The situation is made considerably worse by the increase in viral illnesses like Hepatitis B and C as well as the restricted availability of medical facilities. In order to lower mortality and give patients appropriate treatment, it is necessary to diagnose liver illnesses. But the problem is treatments are not advanced in most of the rural areas of Bangladesh for which it's not possible detect disease early and accurately. To solve these problems, we propose a machine learning-based system for liver disease prediction. We used clinical data, including biochemical and demographic features as inputs. After necessary pre-processing, the dataset was trained using various machine learning algorithms, including Decision Tree, Random Forest, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), XGBoost, LightGBM, CatBoost, and Explainable Boosting Machines (EBM). Among these, EBM achieved the highest test accuracy of 99.72% after hyperparameter tuning.
Index Terms-Dataset, Disease detection, machine learning, liver disease, supervised learning.

1.Introduction

As one of the most densely populated countries in the world, Bangladesh faces numerous public health challenges. Non-communicable diseases, such as liver disease, have become a significant concern due to factors like the high prevalence of Hepatitis B and C infections and limited healthcare access in rural areas [1]. Liver diseases not only impact individuals' health but also place a substantial economic burden on the healthcare system. [1]

Liver disease is a leading cause of morbidity and mortality worldwide, responsible for millions of deaths each year. In Bangladesh, liver diseases are a major public health concern. Due to liver complications from chronic infections, hepatisis B and C each kill 300000 to 800000 people a year globally [2]. Hepatitis B is the most common liver infection in the world, but fewer than 10% of persons with hepatitis B are diagnosed, and fewer than 2% of persons eligible for

treatment receive lifesaving medications [3]. In Bangladesh Chronic liver disease are the most common form of liver disease. Hepatic Encephalopathy is a complication of Chronic liver disease. In regions like Dhaka, where healthcare is relatively better, only 2.6% of CLD patients develop HE. In regions like Khulna, with less advanced healthcare systems, the prevalence of HE rises to 13.6% [4]. It clearly visible that there is a huge demand for a disease detecting system that needs less resources. So we think building a machine learning based disease detection system can reduce the dependencies on the old manual disease detection method and will greatly benefit the people from rural areas to get faster and more accurate results [2].

To detect liver disease accurately and efficiently we built a machine learning-based system that analyzes clinical biomarkers like bilirubin and albumin for precise predictions. We've also used advanced preprocessing techniques including normalization and oversampling, ensure data quality and also optimized algorithms like XGBoost, LightGBM, and EBM enhance computational efficiency which makes the system ideal for less resourced environments. What sets us apart is that, it focuses on using EBM to help healthcare professionals understand how predictions are made and compares multiple algorithms to determine the most effective model. This ensures a scalable and transparent diagnostic solution which is very suitable for low resourced areas of Bangladesh. [5] [6] [7]

In this paper our aim is to investigate how some clinical and biochemical data can contribute to predict liver disease of people of Bangladesh. By analyzing clinical datasets that contains biomarkers such as bilirubin, albumin, and alkaline phosphatase. With these data's, we aim to build a system that can predict liver disease with high accuracy. [8] Our research seeks to answer the following questions:

1. Can machine learning models trained on publicly available datasets effectively predict liver diseases in Bangladeshi patients?
2. How do advanced machine learning algorithms, including Explainable Boosting Machines (EBM) and XGBoost, compare in terms of accuracy and computational efficiency for liver disease prediction?

As the resources in Bangladesh are very limited, we aimed to show that with publicly available data set like the one we used [8] can be utilized to train learning problems applicable to the local people of Bangladesh. This could solve the problem of dataset scarcity which can leave a huge

impact on medical sector of Bangladesh. The insights from this study will help healthcare professionals understand how certain clinical factors and biomarkers are related to liver disease. This knowledge will support better decision-making and ensure patients receive treatment at a much quicker and accurate manner.

We trained multiple machine learning models for liver disease prediction using a publicly available clinical dataset [8]. Our contributions reported in this article are primarily two-fold:

- **Data Quality and Preprocessing**:
  During our experiments, we identified and addressed several data inconsistencies such as missing values and imbalanced classes, which could hinder model performance. These issues were corrected through advanced preprocessing techniques, such as data normalization and oversampling, ensuring that the dataset was well-prepared for accurate predictions. This approach to data handling has been crucial in achieving high performance and reliability, distinguishing our work from previous studies that did not implement such comprehensive preprocessing methods. [9]
- **High Accuracy and Computational Efficiency:**
  Unlike previous works that focused primarily on model accuracy without considering computational efficiency, our models achieved not only high accuracy (99.72%) but also significantly reduced computational costs. For instance, our models were trained and tested with minimal computational cost, making them highly efficient for real-world deployment in resource constrained environments. This combination of high accuracy and computational efficiency makes our system more scalable and practical than existing models. [10] [5] [6] [11]

These contributions highlight the potential of machine learning to improve liver disease diagnosis, offering a practical, accurate, and efficient solution. By focusing on both model accuracy and computational efficiency we aim to provide a scalable solution that can be effectively utilized in regions with limited healthcare resources.

The rest of the paper organized as follows: In section 2 we reviewed some other works done by different people and organization and compared them. Then on section 3 which is titled methodology, we described the process of the whole work. In section 4 we showed the results we got after testing our model with multiple algorithms. In section 5 we compared our models performance with some other works that were conducted on this topic and lastly section 6

contains concluding remark and our future plan regarding this project.

2. Literature Review:

Recent studies have highlighted the growing potential of machine learning (ML) in the prediction and classification of liver diseases. One such study focused on the Random Forest (RF) algorithm which achieved an accuracy of 88%. The study emphasized the importance of feature selection methods, such as Particle Swarm Optimization, to optimize the model's performance. Additionally, data preprocessing techniques, including handling missing values and addressing data anomalies, were pivotal in enhancing the model's accuracy. The use of performance metrics such as precision, recall, and confusion matrices further validated the model. The research concluded that RF was highly effective for liver disease classification and recommended exploring hybrid models to improve predictive accuracy further. [10]

In another study, Autoencoders and K-Nearest Neighbors (KNN) were evaluated for liver disease detection. Autoencoders achieved an accuracy of 92.1%, with KNN closely following at 91.7%. The study utilized the ROSE technique to balance the dataset, addressing issues of data imbalance and missing values. A detailed correlation analysis revealed that bilirubin levels were critical predictors, while the albumin-to-globulin ratio had less predictive power. These findings highlighted the importance of selecting relevant features for disease detection. The study concluded that Autoencoders and KNN are reliable methods for early liver disease detection and suggested integrating them into clinical workflows to enhance diagnostic accuracy. [5]

Further advancements were made by integrating feature extraction techniques such as Principal Component Analysis (PCA), Factor Analysis (FA), and Linear Discriminant Analysis (LDA) into liver disease prediction. These techniques led to significant improvements in accuracy, with Random Forest achieving an accuracy of 88.1% and an AUC score of 88.2%. The study found that combining these feature extraction methods improved the model's ability to capture complex patterns in the data. It also demonstrated the effectiveness of integrated feature extraction and benchmarking, suggesting that ensemble models could be explored to boost prediction accuracy even further. [6]

Ensemble learning approaches were another focus of research, with Gradient Boosting (GB) emerging as the top performer, achieving an impressive accuracy of 98.8%. XGBoost, which excelled in recall and AUC (0.987), was also highlighted as a strong contender. The study used techniques like SMOTE for data balancing, MinMaxScaler for normalization, and imputation for missing values to preprocess the data. The research found that ensemble learning methods significantly outperformed traditional machine learning algorithms, providing more robust solutions for liver disease prediction. While the study acknowledged the trade-offs between model complexity and runtime efficiency, it recommended exploring hybrid ensemble methods and incorporating additional biomarkers to improve predictive capabilities. [11]

Table: Notable work done by other people

Report Title	Publication	Key Findings	Algorithms Used	Performance	Techniques Highlighted
Liver Disease Prediction and Classification using Machine Learning Techniques [10]	*International Journal of Advanced Computer Science and Applications (IJACSA), 2023*	Random Forest (RF) outperformed KNN and Logistic Regression, highlighting feature engineering.	Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression	88% accuracy (RF)	Feature selection (Particle Swarm Optimization), data preprocessing (missing values, anomalies).
Liver Disease Detection Using Machine Learning Techniques [5]	*CITRENZ 2022 Conference*	Autoencoders and KNN were effective for liver disease detection, with data balancing.	Autoencoders, K-Nearest Neighbors (KNN)	Autoencoders 92.1%, KNN 91.7%	ROSE technique for data balancing, feature relevance analysis.
Prediction of Chronic Liver Disease Patients Using Integrated Projection-Based Statistical Feature Extraction with	*Informatics in Medicine Unlocked, 2023*	PCA, FA, and LDA improved accuracy, with Random Forest leading the performance.	Random Forest (RF)	88.1% accuracy, AUC 88.2%	Integrated feature extraction (PCA, FA, LDA), benchmarking

Machine Learning Algorithms [6]					
Improved Liver Disease Prediction Using Ensemble Learning Approaches [11]	*BMC Medical Informatics and Decision Making, 2024*	Gradient Boosting (98.8%) outperformed other models, with XGBoost excelling in recall and AUC.	Gradient Boosting (GB), XGBoost	98.8% accuracy (GB), AUC 0.987 (XGBoost)	SMOTE (data balancing), MinMaxScaler (normalization), imputation, hybrid ensemble methods

3. Methodology

Block Diagram:

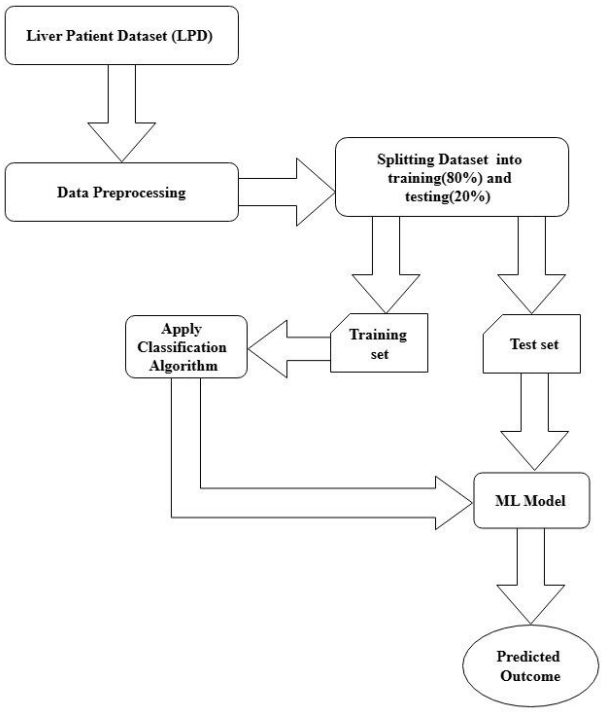

Figure: Block Diagram of the entire work

Dataset

The liver disease dataset was created by Abhishek Shrivastava. This data set contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. The target variable, "Result" shows whether a person has liver disease or not.To improve the accuracy and quality we had to remove duplicate data. Before removing the size of the data was 30691 and after removing duplicates value the size of the data was 19368. [8]

Data Preprocessing

The dataset was then split into training and test sets in an 80:20 ratio, resulting in 15,494 instances in the training set and 3,874 instances in the test set. Before splitting, missing values in the dataset were addressed. For each feature, missing values in the training data were replaced with the feature's mean value. These computed means were also applied to the test dataset to maintain consistency and prevent data leakage. To standardize the scale of the features, a Min-Max Scaler was used, normalizing all attribute values to a range between 0 and 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Pearson's Correlation Coefficient was used for feature selection in order to find and remove strongly correlated characteristics. A threshold of 0.85% was established.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

From this process we've developed a heatmap of correlated features:

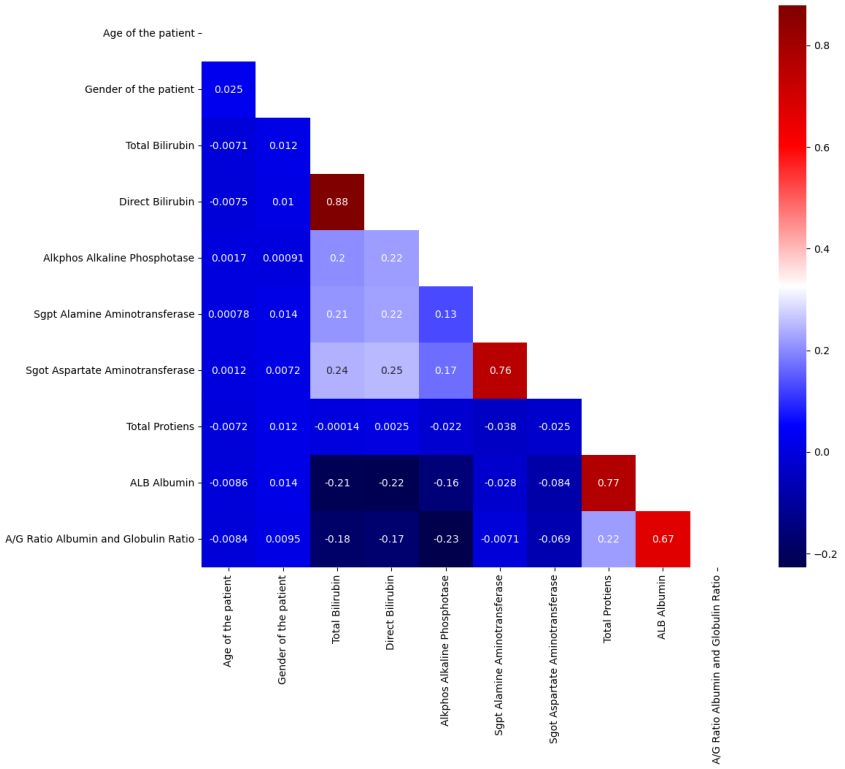

Figure: Correlation Heatmap

During this process, Direct Bilirubin was identified as highly correlated with Total Bilirubin. So we dropped that to improve our accuracy of model.

Model Training

The processed dataset was used to train multiple classification algorithms. Each algorithm consist of different methods and provides different accuracy. We tested every algorithm to find out which performs the best

Decision Tree Classifier:

The Decision Tree model works by recursively splitting the dataset into smaller subsets based on feature values. Each split is made to reduce the impurity in the data, using entropy or information gain to determine the best feature to split on. The tree continues to grow until it reaches a predefined maximum depth or the data in the node is sufficiently pure:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

$$IG(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

This model is simple to interpret, as it creates a clear set of decision rules. The training accuracy was 99.99%, and the test accuracy was 98.84%.

Random Forest Classifier:

Random Forest is an ensemble learning method that creates multiple decision trees and combines their predictions to improve performance. Each tree is trained on a random subset of the data, and the final prediction is determined by majority voting:

$$\hat{y} = \text{Mode}(T_1(x), T_2(x), \ldots, T_n(x))$$

This approach helps reduce overfitting by averaging the predictions of several trees, leading to a more robust model. The training accuracy was 99.99%, and the test accuracy was 99.35%.

Gradient Boosting Classifier:

Gradient Boosting builds an ensemble of trees in a sequential manner, where each new tree tries to correct the errors made by the previous trees. The model is updated iteratively by minimizing the loss function, such as log-loss for classification tasks:

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x)$$

This model works well with complex datasets and is effective in capturing non-linear relationships. The training accuracy was 88.85%, and the test accuracy was 88.31%.

k-Nearest Neighbors (k-NN):

The k-NN algorithm predicts the class of a new instance based on the majority class of its *k* nearest neighbors in the training set. The Manhattan distance (also known as L1 distance) between data points is used to measure proximity:

$$d(x_i, x_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

The value of *k* was tuned using cross-validation to determine the optimal number of neighbors for classification. The training accuracy was 89.78%, and the test accuracy was 81.80%.

Support Vector Machine (SVM):

SVM aims to find the optimal hyperplane that best separates the classes in the feature space. It maximizes the margin between the classes to improve generalization:

$$w \cdot x + b = 0$$

The model was tuned with different kernel functions (such as RBF) to handle non-linear relationships between the features. The training accuracy was 71.36%, and the test accuracy was 71.12%.

LightGBM:

LightGBM is a gradient boosting framework designed for speed and efficiency. It uses a leaf-wise growth strategy for decision trees, which allows it to model complex relationships more effectively than level-wise methods:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) - \gamma$$

This approach ensures fast training and is particularly effective for large datasets. The training accuracy was 99.75%, and the test accuracy was 99.33%.

CatBoost:

CatBoost is another gradient boosting model that excels at handling categorical features. It uses target encoding to convert categorical variables into numerical values:

$$\text{Encoded Value} = \frac{\sum_{j \in S} y_j + a \cdot \text{Prior}}{|S| + a}$$

This model is highly efficient and reduces the need for extensive preprocessing of categorical features. The training accuracy was 99.93%, and the test accuracy was 99.30%.

Explainable Boosting Machine (EBM):

EBM is a machine learning algorithm designed to provide high accuracy while being interpretable. It is a form of generalized additive model (GAM), where each feature's contribution to the prediction is explained in a transparent manner. The model works by building individual additive models for each feature and combining them for a final prediction. It effectively balances performance with interpretability:

$$F(x) = \sum_{j=1}^{p} f_j(x_j)$$

EBM focuses on interpretability and provides insight into how individual features influence the final prediction. The training accuracy was 99.96%, and the test accuracy was 99.66%.

Evaluation Metrics

The performance of the models was evaluated using a range of metrics to ensure their predictive power and reliability:

Accuracy:

Accuracy measures the proportion of correct predictions out of all predictions:

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

where TP, TN, FP, and FN represent the True Positives, True Negatives, False Positives, and False Negatives, respectively.

Precision:

Precision evaluates the proportion of correctly predicted positive instances out of all predicted positives:

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

Recall:

Recall measures the ability of the model to correctly identify all positive instances:

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

F1 Score:

The F1 score is the harmonic mean of precision and recall, providing a balanced measure of model performance:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Then we applied hyper parameter tuning to get more accuracy out of these models. Among all the algorithms the Explainable Boosting Machine (EBM) achieved the highest accuracy, demonstrating its effectiveness for this dataset. The structured approach to preprocessing, feature selection, and model tuning ensured reliable and interpretable results, showcasing the potential of machine learning in diagnosing liver diseases.

4. Result

Table: Train and test accuracy of different models:

Model	Train Accuracy (%)	Test Accuracy (%)
Decision Tree	99.99%	98.84%
Random Forest	99.99%	99.35%
k-Nearest Neighbors	89.78%	81.80%
Support Vector Machines (SVM)	71.36%	71.12%
Gradient Boosting Machines (GBM)	88.85%	88.31%
XGBoost	99.98%	99.43%
LightGBM	99.75%	99.33%
CatBoost	99.93%	99.30%
Explainable Boosting Machine (EBM)	99.96%	99.66%

Here we calculated train and test accuracy with different algorithms. As it seems that EBM provides the highest accuracy amongst all the other algorithms. For even better accuracy we've decided to apply hyper parameter tuning.

Table: Train and test accuracy of different models after hyper parameter tuning:

Model	Best Parameters	Train Accuracy (%)	Test Accuracy (%)
Decision Tree	`{'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': None, 'criterion': 'gini'}`	99.99%	98.84%
Random Forest	`{'n_estimators': 300, 'min_samples_split':	99.97%	99.46%

	10, 'min_samples_leaf': 1, 'max_depth': 20}`		
k-Nearest Neighbors	`{'weights': 'distance', 'n_neighbors': 3, 'metric': 'manhattan'}`	99.99%	93.42%
Support Vector Machines (SVM)	`{'kernel': 'poly', 'gamma': 'scale', 'C': 100}`	71.69%	71.61%
Gradient Boosting Machines (GBM)	`{'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.1}`	99.99%	99.43%
XGBoost	`{'subsample': 1.0, 'n_estimators': 300, 'max_depth': 5, 'learning_rate': 0.2, 'colsample_bytree': 0.8}`	99.99%	99.51%
LightGBM	`{'num_leaves': 50, 'n_estimators': 200, 'max_depth': -1, 'learning_rate': 0.1}`	99.99%	99.56%
CatBoost	`{'learning_rate': 0.2, 'iterations': 300, 'depth': 6}`	99.99%	99.33%
Explainable Boosting Machine (EBM)	`{'min_samples_leaf': 1, 'max_leaves': 3, 'max_interaction_bins': 32, 'max_bins': 512, 'learning_rate': 0.1, 'interactions': 5}`	99.94%	99.72%

Hyperparameter tuning significantly improved the performance of most models by improving test accuracy and reducing overfitting. As we can see models like k-Nearest Neighbors (k-NN) showed the greatest improvement as its test accuracy went from 81.80% to 93.42% due to optimized parameters like n_neighbors and weights. The other models that performed well like Random Forest, Gradient Boosting Machines (GBM), XGBoost, and LightGBM also experienced modest accuracy gains and improves reliability.

The Explainable Boosting Machine (EBM) emerged as the best-performing model, achieving a test accuracy of 99.72% after fine-tuning.

5. Discussion:

The results that we achieved on our project demonstrates the high possibility of machine learning algorithms in liver disease detection. On our testing, expandable boosting machine (EBM) achieved the highest accuracy of 99.72%. To achieve this number we used some notable techniques such as hyperparameter tuning, imputation and Pearson's correlation for feature selection. All those altogether improved the overall precision, recall and accuracy of the model.

Table: Comparison of our model with other notable works

Report Title	Key Findings	Algorithms Used	Performance	Techniques Highlighted
Liver Disease Prediction and Classification using Machine Learning Techniques [10]	Random Forest (RF) outperformed KNN and Logistic Regression, highlighting feature engineering.	Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression	88% accuracy (RF)	Feature selection (Particle Swarm Optimization), data preprocessing (missing values, anomalies).
Liver Disease Detection Using Machine Learning Techniques [5]	Autoencoders and KNN were effective for liver disease detection, with data balancing.	Autoencoders, K-Nearest Neighbors (KNN)	Autoencoders 92.1%, KNN 91.7%	ROSE technique for data balancing, feature relevance analysis.
Prediction of Chronic Liver Disease Patients Using Integrated Projection-Based Statistical Feature Extraction with Machine Learning Algorithms [6]	PCA, FA, and LDA improved accuracy, with Random Forest leading the performance.	Random Forest (RF)	88.1% accuracy, AUC 88.2%	Integrated feature extraction (PCA, FA, LDA), benchmarking
Improved Liver Disease Prediction Using Ensemble Learning Approaches [11]	Gradient Boosting (98.8%) outperformed other models, with XGBoost	Gradient Boosting (GB), XGBoost	98.8% accuracy (GB), AUC 0.987 (XGBoost)	SMOTE (data balancing), MinMaxScaler (normalization), imputation, hybrid ensemble methods

	excelling in recall and AUC.			
A Comprehensive Analysis of Liver Disease Detection Using Advanced Machine Learning Algorithms (Proposed Work)	Explainable Boosting Machine (EBM)(99.72%) performed best among all and only XGBoost(99.51%) and LightGBM(99.56) came close.	Decision Tree, Random Forest, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), XGBoost, LightGBM, CatBoost, and Explainable Boosting Machines (EBM)	99.72% accuracy (EBM)	Mean Imputation(MI), MinMaxScaler(normalization),Feature selection Using Pearson's correlation

Our research represents significant improvement when compared to previous studies. While earlier works on liver disease detection achieved strong results using models like Random Forest, Gradient Boosting, and XGBoost with accuracy levels reaching up to 98.8% [11]. With our approach we can reach beyond that level of accuracy. By utilizing the Explainable Boosting Machine (EBM) we achieved better accuracy. We also prioritized model interpretability which makes our work both precise and transparent. Finally we incorporated advanced techniques like MinMaxScaler, Feature selection Using Pearson's correlation and mean imputation. These techniques ensured that our model is capable of handling imbalanced datasets and noisy features with exceptional robustness which sets our project apart from others.

6. Conclusion

This paper presents a machine learning-based approach for liver disease detection. The work holds significant importance in improving healthcare outcomes of Bangladesh. Specially detecting liver disease at an early stage which will greatly benefit the people from rural. A comparative analysis of eight machine learning algorithms, including Decision Tree, Random Forest, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), XGBoost, LightGBM, CatBoost, and Explainable Boosting Machine (EBM), was conducted. The models predicted liver disease with different kinds of accuracy. Amongst everything EBM achieving the best performance at 99.72% accuracy on test data.
As we have identified a near-optimal algorithm, we plan to extend this work further by utilizing more comprehensive and higher-quality datasets in the future. We would also like to integrate clinical expert feedback in future to enhance the system's practicality and effectiveness in real-world applications.

References

[1] "Hepatitis B," world health organization, 092024 April 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b.

[2] J. Trefethen, "Open Philanthropy," COALITION FOR GLOBAL HEPATITIS ELIMINATION, 2024. [Online]. Available: https://www.globalhep.org/. [Accessed 2024].

[3] D. J. W. Ward, "COALITION FOR GLOBAL HEPATITIS ELIMINATION 5 YEARS TOGETHER, ELIMINATING HEPATITIS," 2024.

[4] M. F. A. ,. M. J. A. c. R. D. M. I. H. M. M. H. A. S. K. M. F. K. Salimur Rahman, "Distribution of Liver Disease in Bangladesh: A Cross-country Study," 22 January 2014.

[5] D. C. N.-L. T. S. S. T. S. K. R. Bhupathi, "Liver disease detection using machine learning techniques," 2022.

[6] R. Y. S. R. ,. M. H. R. M. S. R. Ruhul Amin, "Prediction of chronic liver disease patients using integrated projection".

[7] S. S. A. Nilofer, "A Comparative Study of Machine Learning Algorithms Using Explainable Artificial Intelligence System for Predicting Liver Disease," World Scientific, Chennai, Tamil Nadu, India, 2024.

[8] A. Shrivastava, "Liver Disease Patient Dataset 30K train data," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset.

[9] A. H. A. H. K. T. A. L. Abroor Zahin Niloy, "cleaned_dataset for Liver Patient Dataset (LPD)_train.csv," [Online]. Available: https://drive.google.com/file/d/1eI5NtwOhqXwuSNyca1khVkOdk51XOeVS/view.

[10] K. H. R. P. G. B. N. N. A. K. E. Srilatha Tokala, "Liver Disease Prediction and Classification using Machine Learning Techniques," (IJACSA) International Journal of Advanced Computer Science and Applications, SRM University-AP, Amaravati, India, 2023.

[11] P. K. D. P. Z. Shahid Mohammad Ganie, "Improved Liver Disease Prediction from Clinical Data Through an Evaluation of Ensemble Learning Approaches," worldscientific, Chennai, Tamil Nadu 600 005, India, 2024.