RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2553

Pre-Symptomatic Liver Disease Identification Using Distinct Machine Learning Methods

Sumathi Selveraj [a, *], Ramesh Thenappan [a], Parthiban Chakkaravarthi [b]

[a] Department of Information Technology, Bharathiar University, Coimbatore-641046, Tamil Nadu, India.

[b] Department of Computer Science, Indo-American college, Cheyyar-604407, Tamil Nadu, India

* Corresponding Author Email: sumathiparthi88@gmail.com

DOI: https://doi.org/10.54392/irjmt2553

Received: 17-02-2025; Revised: 30-07-2025; Accepted: 23-08-2025; Published: 12-09-2025

Abstract: The liver which serves as a crucial human organ sustains metabolic operations while preserving health levels across the body. Successful treatment along with improved patient results require both quick and precise detection of liver disease through diagnosis. This study uses Machine Learning (ML) algorithms including Random Forest (RF), Categorical Boosting (CB), Adaptive Boosting (AB), Light Gradient Boosting Machine (LGBM) and Support Vector Classification (SVC) together with Logistic Regression (LR) to establish a method for Liver Disease (LD) prediction. Our research investigated numerous methods to assess their successful classification ability for liver disease with accuracy rate evaluations alongside precision and recall statistics and F1 score calculations. Notably, to make the reliability and generalizability of each model, the cross-validation was carried out. Our multi-algorithmic strategy boosts prediction stability by providing both precise analysis of each algorithm's predictive capabilities and their respective weaknesses in liver disease forecasting. CB due to its high accuracy and robustness, has produced the best results in comparison with other algorithms, which were analyzed. The ability to effectively deal with categorical variables and the low level of intensive preprocessing greatly contributed to its better performance in comparison with conventional models.

Keywords: Liver Disease, Disease Prediction, Machine Learning, Algorithm, Classifying

1. Introduction

Two million individuals throughout the world lose their lives to liver disease annually with half dying from circulatory complications associated with cirrhosis and the other half succumbing to hepatocellular carcinoma and viral hepatitis [1]. As the human body's essential organ the liver executes various metabolic functions while simultaneously protecting overall bodily health. The liver performs these three functions: waste filtering alongside both energy conservation and red blood cell breakdown. The liver removes blood coming from digestion before performing disinfection functions. The liver performs dual functions by cleaning drugs along with dissolving toxic chemicals. Through its internal processes the liver generates proteins required by the body to coagulate blood. The liver extracts foreign substances and infections which enter the bloodstream. Untreated liver disorders lead to permanent damage while simultaneously compromising functional liver health. Due to sympathecious outer presentation liver disease detection proves challenging during the early stages of liver health deterioration. The effective accessing of needed treatments through medication demands patients to receive both early detection and accurate diagnoses [2]. Global statistics show liver diseases stand among the top death causes since their numbers grew considerably while their severity increased during recent decades. According to the World Health Organization (WHO) chronic conditions result in yearly fatalities of 35 million people and comprise 46% of known diseases while masking 59% of total worldwide deaths [3]. The application of ML shows promises to boost objective decision-making while diagnosing and forecasting diseases which biomedical researchers find particularly important [4]. ML techniques enable quick diagnosis solutions while also minimizing diagnostic costs. Outcomes of this project focus on enhancing prognostic predictions while reducing healthcare diagnostic expenses. The study team used several categorization methods to identify whether patients had liver disease. Directional performance analysis of the RF, CB, XGB, LGBM, SVC and LR ML techniques checked their accuracy while measuring precision, recall and f score. In our analysis, we used the most commonly used evaluation metrics such as precision, recall, F1-score, to compare the performance of multiple categorisation models. To assess the discrimination capability of each model further we have computed the Receiver Operating

Characteristic (ROC) curve and the Area Under the Curve (AUC) values. These indices provided a complete understanding of the performance of each model, more so in relation to class separation of liver disease prediction. The same kind of methodology was applied to evaluate the model performance by ROC-based analysis, which is reported in [5].

Medical diagnostic modalities including magnetic resonance imaging (MRI) and computed tomography (CT) along with ultrasound (US) serve as essential tools for detecting liver illness [6, 7]. Global implementation of ML by researchers continues to grow as this technology attracts broad attention from researchers around the world. Computer science technology called ML proves successful across multiple crucial medical conditions including cardiovascular issues and cancer treatments and dengue fever diagnostics [8]. The problem of excessive dimensions in ML data surfaces frequently during operations. The large amounts of required memory while operating from high-dimensional datasets lead to overfitting risks through redundant and unnecessary data instances. Feature selection becomes necessary to handle this situation. The designronym for selecting useful features while discarding unneeded ones is called "feature selection" [9]. Doctors require multiple tests to check liver health but these diagnostic methods are both lengthy and complicated. ML solutions enable the assessment of predictions to bypass delicate and costly examination procedures. Through early detection of liver diseases and potential disease progression control the requirement for invasive examinations and complex medical treatments might vanish completely [10]. Lots of research has explored different ML algorithms as potential methods to predict LD. Additional research needs to determine different hyper parameter settings of ML algorithms to boost model accuracy while improving dependability. To enhance performance of models, this paper proposes a prediction model of liver diseases (LD) based on ML. The use of ML refers to a set of computational techniques that allow systems to learn patterns in data and predicts outcomes without additional programming. Utilizing the ML techniques, the proposed approach will be able to analyses the complex medical information, identify valuable traits and accurately predict the progression of hepatic disease development. This approach is essential to make clinical decisions, diagnose early, and enhance predictive accuracy. This system seeks to detect potential health threats and establish early warning capabilities and determine ideal hyper-parameter configurations [11].

Most of the publishable studies in machine learning in the field of liver disease prediction done in recent years (2020-2025) focused on optimizing the accuracy of a model by exploring the tuning of the algorithm, preprocessing technique. Although such models are generally effective on a typical dataset, they do not often account for such clinically relevant factors as age-specific risk factors, interpretability to practical application, and demographic fairness. Our method addresses this topic by ensuring the stability of performance and the overall applicability to practice by relying on cross-validation and operating a unity approach of a variety of classifiers, including CB, RF, LGBM and AB. Compared with other studies, very few factors are looked at in a comprehensive manner such as subgroup assessment, lifestyle features (e.g., alcohol intake, BMI, smoking, physical exercise), and stratified age categories that we examine partly due to the absence of parametric statistical tests, which prohibit their consideration in the majority of cases. Moreover, to the best of our knowledge, statistical validation, and cross-comparison of models along multiple metrics (F1, AUC, precision, recall) is an objective that has not been achieved in the literature because our work extends beyond accuracy maximisation to develop reliable, explanable, and demographically aware liver disease prediction models that can be used in clinical settings.

2. Related Works

Machine learning has been employed in various ways through recent research whereby different methods of liver disease prediction have been discussed. Mostafa *et al*. [12] used RF, Artificial Neural Network (ANN), and Support Vector Machine (SVM) classifiers, as well as preceding the principal component analysis (PCA) that reduced the dimensionality and multiple imputations through chained equations. Their preprocessing involved Gini index relevance ranking in ranking of features and SMOTE to deal with imbalances in classes. RF was the most successful model post parameter tuning, which consists of forest size optimization. The ANN training was performed by 32 and 30 epochs and a range of batch sizes, whereas, SVM used radial basis function kernel with optimal parameters of tuning set to 10. In an experiment, a similar concept was presented by Amin *et al.* [13], mixing three feature engineering methods with the feature-engineering techniques of six classification methods. Their approach to outliers consisted of their natural elimination, their modal replacement based on missing values and their multi-feature fusion. This strategy spent 88.10% of its time on accuracy, 92.30% on recall, 85.33 on precision, F1 score of 88.68, and 88.20 on AUC with 75/25 stratified train-test spilt [14].Joloudari *et al.* [15] have performed a comparison between five data mining models as; RF, multilayer perceptron neural networks (MLP), Bayesian networks, SVM, and a PSO-optimized SVM (PSO-SVM). The experiment done by them and supported by the 10-fold cross-validation provided that best performance was achieved by PSO-SVM on ILPD dataset. Another extensive contribution was realized by Ganie and Dutta Pramanik [16], which developed GB based single or multi-output-based framework for predicting chronic liver disease (CLD) and compared it with AdaBoost, LogitBoost, SGBoost, XGBoost, LGBM, and CB.

Table 1. Compare Recent Methods

Research	Disease	Models Used	Discussion	Year
Wang, Y, *et al* [17]	NAFLD	RF, XGB, SVM	Diagnostic prediction model for NAFLD	2025
Saleem, [18]	Hepatitis C	CNN, SVM, RF	Machine Learning Classification Algorithms	2024
Wang M, *et al* [19]	HCC	Multi-omics + ML	Improved prognosis using omics data	2024
Zhu, H.,*et al* [20]	Liver Metastasis	XGB, SHAP	Interpretable liver metastasis risk model	2025
Wang, X *et al* [21]	NAFLD/CHD	XGBoost, LGBM	Immuno-inflammatory index prediction	2025
Sheakh, *et al* [22]	Hepatitis C	RF	Improving hepatits C diagnosis	2024
Song, Q *et al* [23]	Trauma	DL Ultrasound	Early trauma detection in liver	2025
Wu, Y., *et al* [24]	Liver ADRs	RF, SVM, XGB	Prediction of adverse drug reactions	2025

Out of all the other models used, Gradient Boosting offered the best accuracy of 98.80% on the Liver Disease Patient Dataset (LDPD) and 98.29% on the Indian Liver Patient Dataset (ILPD). All these results are an indication of the increasing significance of ensemble and boosting algorithms like RF, XGBoost and LightGBM in ML applications in the field of hepatology.

Table 1 compares recent methods based on machine learning techniques in diagnosing, prognosis, and classification of liver diseases. The table enumerates the major contributions, ML models, years of publication and focus of the disease. XGBoost, Random Forest, and numerous other algorithms are presented to prove the increased importance of artificial intelligence in the research of hepatology, including other deep learning architectures.

3. Resource and Techniques

3.1 Data Gathering

The data recorded in the current research was downloaded through one of the Kaggle repositories [25] called Predict the Liver Disease: 1700 Records Data set by the author Rabie El Kharoua. It is available to the general population at: https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset with the Creative Commons Attribution 4.0 International License (CC BY 4.0). The data includes 11 attribute datasets that include demographic attributes (e.g. age, gender) and clinical attributes (e.g. total bilirubin, ALT, AST, albumin) concerning 1,700 records of patients. The target Y is categorical (present or absent livers disease). The distribution of the classes is slightly unbalanced: 1,046 cases (61.5%) belong to the patients with the liver disease, and 654 cases (38.5%) are the patients without the liver disease. This imbalance has been mitigated in the developing of the models with the correct method of resampling and validating such as stratified sampling and SMOTE where such is needed. It was quite even in regard to gender split (50.4 percent men and 49.6 percent women). This dataset will be the basis of our ML models training and testing with an adequate sufficient representation, and mitigated gender bias in prediction. It is because stratified sampling, as well as cross-validation methods, were used, in order to examine the validity of the models, despite the lack of other external data [26].

Figure 1 show the research identified four patient age groups which included middle-aged patients aged 58 on average and senior citizens at 78 while young adults were 25 years old and adults comprised 35-year-old patients. Unique diagnosis frequencies of different illnesses were determined for each age segment to provide disease occurrence details for specific population groups. This examination helps make healthcare decisions and resource management more effective because it considers distinct requirements of different population age groups.

3.2 Diagnosis with age group

Figure 2 shows the diagnosis rates which researchers calculated separately for each age group either with or without liver disease [27]. Statistics demonstrate the prevalence of liver disease increases substantially among people in senior and older age groups. The research data shows Liver Disease exists in 177 Young Adults and 189 Adult groups and 282 Middle-Aged and 274 Senior groups while those without Liver Disease have 216 Young Adult cases and 228 Adult cases and 127 Middle-Aged cases and 174 Senior-age patients. A total of 282 cases of liver disease were detected within the middle-aged group which exceeded the number of senior group cases at 274. The study revealed that adult young people had the least number of liver disease cases at 177. The analysis requires further investigation of fundamental discrepancies between different age groups.

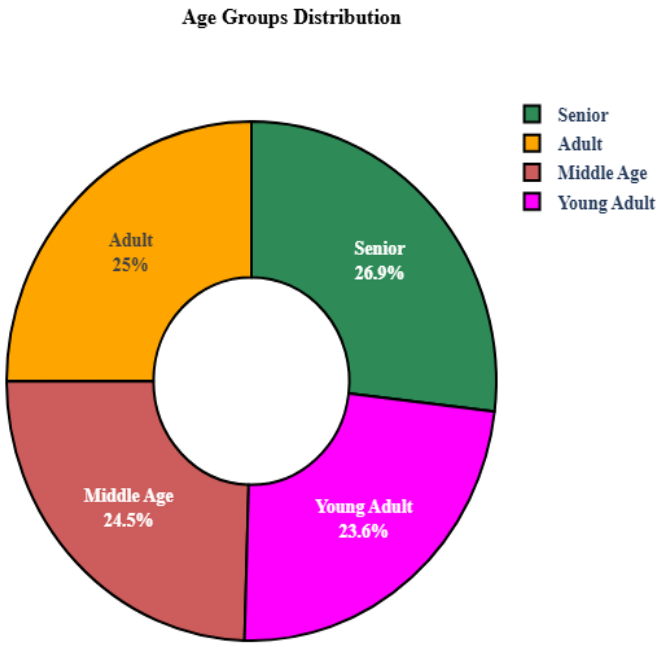

Figure 1. Age Group Distribution

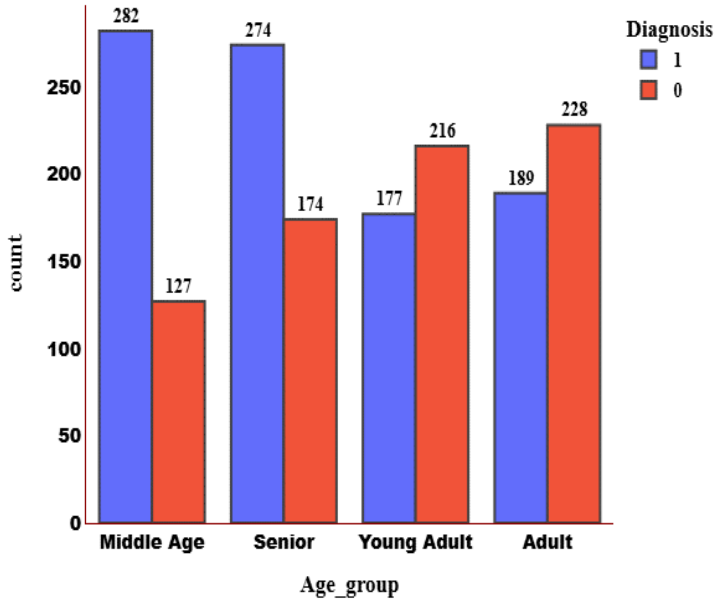

Figure 2. Diagnosis with age group

The medical professional could assess these factors as well as life choices and genetic factors and other illnesses to determine the liver disease risk.

3.3 Alcoholic Consumption and Disease Risk

The Figure 3 demonstrates separate calculations of alcohol units consumed by both liver disease patients and non-liver disease patients within each age category [28]. Research outcomes revealed that patients with liver disease showed an average consumption of 11.95 units among young adults and 7.89 units among the remaining age groups. The group of young adults diagnosed with liver disease averaged 11.95 units per month which was the highest consumption figure 3 among all studied liver disease showed 7.17 to 7.89 units as their average consumption level and remained steady between all age groups. The relationship between two variables should not be mistaken for cause-effect relationships because correlational data alone fails to prove causality. Research must undertake more investigations to prove the direct relationship between alcohol use and liver pathology. Genetic background and lifestyle habits might be extra factors which play a role along with alcohol consumption. The necessary research must involve long-term monitoring of how alcohol use evolves along with liver health assessment through time.

3.4 Exercise hours Vs Liver Functional

Freely accessed data demonstrates the independent analysis of exercise time for people both

with and without liver disease through Figure. 4 Table results show the following data: Without liver disease, Young Adult exercise for 5.49 hours while Adult exercised 5.30 hours, Middle-Aged exercised 5.40 hours, Senior exercised 5.70 hours. When the same groups had Liver Disease, Young Adult exercised 4.50 hours for both groups as did Adult, Middle-Aged exercised 4.89 hours while Senior exercised 4.77 hours. Research indicates that exercise involvement seems to decrease as exposure to liver disease increases [29]. Among all subjects irrespective of their age group people who did not have liver disease commonly exercised more than those with this condition. The exercise duration of 5.70 hours among senior people without liver disease matched the young adult group who exercised 5.49 hours daily. The average exercise hours for people with liver disease amounted to 4.50 to 4.89 hours per year without any consistent differences between age groups.

The reality stands that tight relationships between factors do not necessarily mean those factors create causes to each other. Additional research must be conducted to provide absolute proof that exercise relates to liver health.

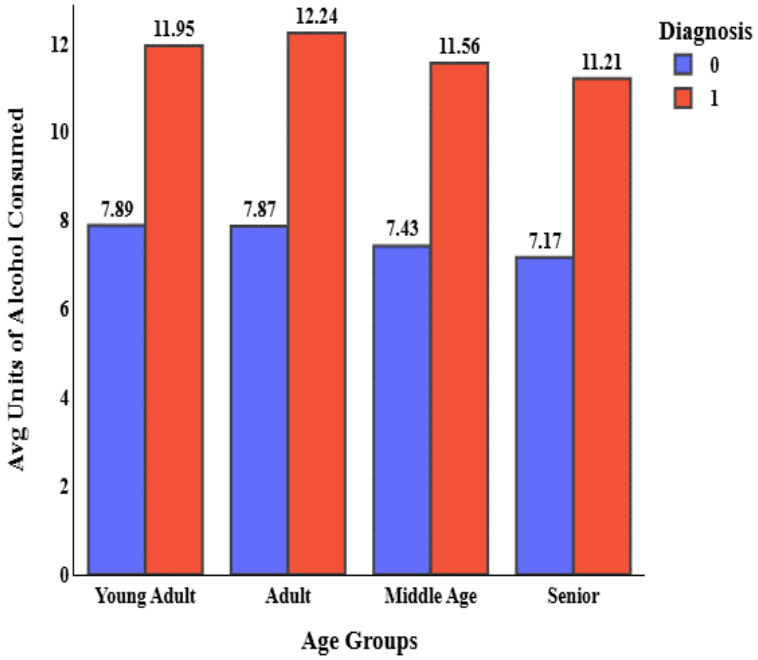

Figure 3. Alcoholic consumption

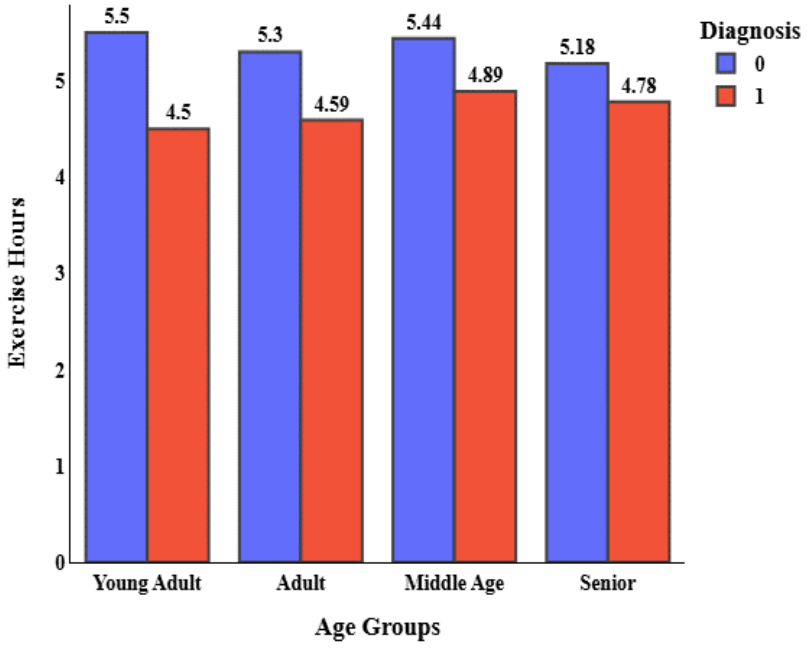

Figure 4. Exercise hour comparison

Diet along with alcohol consumption combined with genetic factors affect how likely a person becomes sick with liver disease. Long-term studies must track both exercise behavior changes and liver health evolution patterns throughout the investigation period.

3.5 Body mass index and a higher risk of liver disease

Figure 5 depicts the BMI average calculations from individual age groups among people who had or did not suffer from liver disease separately [30]. Research findings revealed that Liver Disease participants achieved BMI scores of 27.99 as Young Adults and 28.93 as Adults and 28.75 as Middle-liver disease showed 7.17 to 7.89 units as their average consumption level and remained steady between all age groups. The relationship between two variables should not be mistaken for cause-effect relationships because correlational data alone fails to prove causality. Research must undertake more investigations to prove the direct relationship between alcohol use and liver pathology. Genetic background and lifestyle habits might be extra factors which play a role along with alcohol consumption. The necessary research must involve long-term monitoring of how alcohol use evolves along with liver health assessment through time. Aged individuals yet 29.27 as Seniors while participants free of Liver Disease presented scores of 26.68 Young Adults and 26.57 Adults and 25.30 Middle-Aged along with 26.40 as Senior participants the collected data indicates BMI status might influence liver disease risk rates particularly within the collective age segments of young adults. Among the population with liver disease the youngest (27.99) and oldest adults (29.27) demonstrated the highest average BMIs. Individuals that did not have liver disease showed a constant average BMI reaching between 25 points 30 to 26 points 68 within all age groups. The existence of an association between variables does not necessarily prove that one variable produces effects in the other variables. Additional study is needed to establish a direct link between body mass index (BMI) and liver disease manifestations. Liver disease risk factors include both genetic make-up and alcohol consumption and lifestyle habits. Time-interval scientific studies must be conducted to track BMI fluctuations while monitoring liver health progression.

3.6 Smoking and Disease Risk

Research showed how many people in each age grouping along with diagnostic category currently smoked according to Figure.6. These research findings demonstrate that smoking increases the risk of developing liver disease across every age group because the percentages [31]. Show Young Adult: 0.190; Adult: 0.190; Middle-Aged: 0.180; Senior: 0.180; without liver disease: 0.190; Adult: 0.417; Middle-Aged: 0.366; Senior: 0.306. The proportion of smokers were consistently greater between liver disease patients than between individuals who did not have liver disease throughout all age groups. The young adult group with liver disease demonstrated the highest proportion of smokers at 0.423 while the adult population had the lowest percentage at 0.417. Across all age groups the percentage of smokers among subjects without liver disease displayed minimal variation ranging from 0.180 to 0.190.

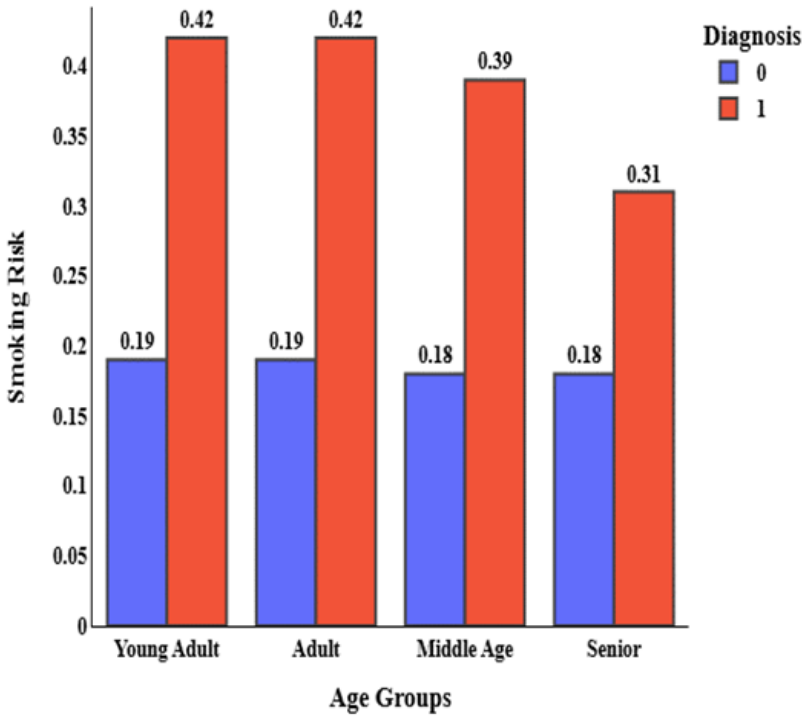

Figure 5. BMI Basic Risk Factor

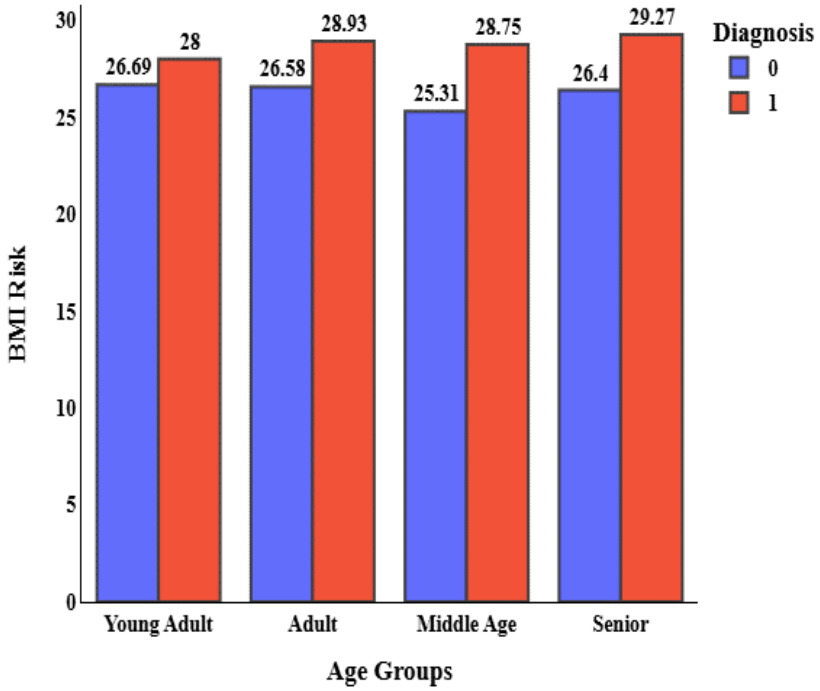

Figure 6. Smoking habit risk factor

The relationship between variables does not necessarily indicate that one thing creates another in cause-effect relationships. Additional studies must be conducted to prove definitively that smoking leads to liver disease development. The risk of developing liver disease depends on how a person lives their life and their alcohol drinking habits and their inherited tendencies in addition to various other factors. Long-term studies must track the development of smoking behaviors along with liver health changes throughout time.

3.7 Diabetes Disease Risk

Researchers calculated the respective percentages of diabetic and non-diabetic individuals in Figure 7 according to their age ranges and medical conditions [32]. Research findings show that diabetes may lead people toward elevated liver disease risk especially within the younger patient groups. The data shows Young Adults held a 0.15 percentage while Adults measured 0.19 and Middle-Aged participants settled at 0.19 and Seniors maintained 0.18 yet Without Liver Disease revealed 0.12 followed by Adults at 0.10 then Middle-Aged at 0.05 and Seniors at 0.11. Among all age groups fewer diabetic persons existed in populations which lacked liver disease compared to people with liver disease. Adults among people with liver disease reported the highest rate of diabetes at 0.18 while those in middle-aged and senior groups shared the same rate of 0.18. The proportion of people without liver disease who maintained diabetes remained steady regardless of their age group spanning from 0.047 to 0.12. We need to keep in mind that when two variables correlate but this does not establish causality. Further investigations about the connection between diabetes and liver disease need to be conducted to confirm their mutual influence. Additional risk factors contributing to liver disease include life choices and alcohol usage together with genetic susceptibility. Longitudinal studies must be used to follow liver health changes as well as diabetes status transformations across time.

3.8 Hypertension Disease Risk

Researchers calculated the percentages of hypertensive and normotensive patients in Figure 8 through age-based category analysis [33]. Without Liver Disease Groups Current Analysis Shows Young Adults at 0.092 and Adults at 0.087 and Middle-Aged Subjects at 0.086 and Officers Senior Citizens at 0.074 whereas Patients with Liver Disease Exhibited 0.250 for Adults and 0.216 for Middle-Aged Respondents along with 0.220 for Senior Participants. The research data demonstrates that hypertension creates greater susceptibility to liver disease particularly among the younger adult population. A significant percentage increase of hypertensive patients existed between those who had and those who did not have liver disease across all tested age groups. The number of hypertensive patients dealing with liver disease reached its peak at 0.25 within the young adult population followed by 0.216 in the adult group. People who did not have liver disease presented a hypertension rate range from 0.074 to 0.092 across age groups which remained relatively stable. The truth remains that the relationship between two variables does not automatically mean they have a cause-effect relationship.

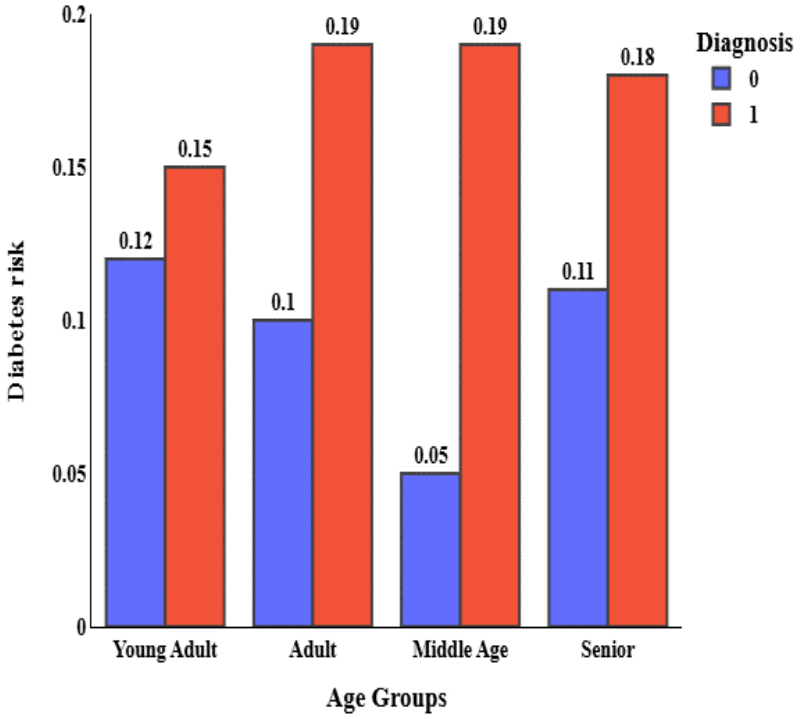

Figure 7. Diabetes Risk

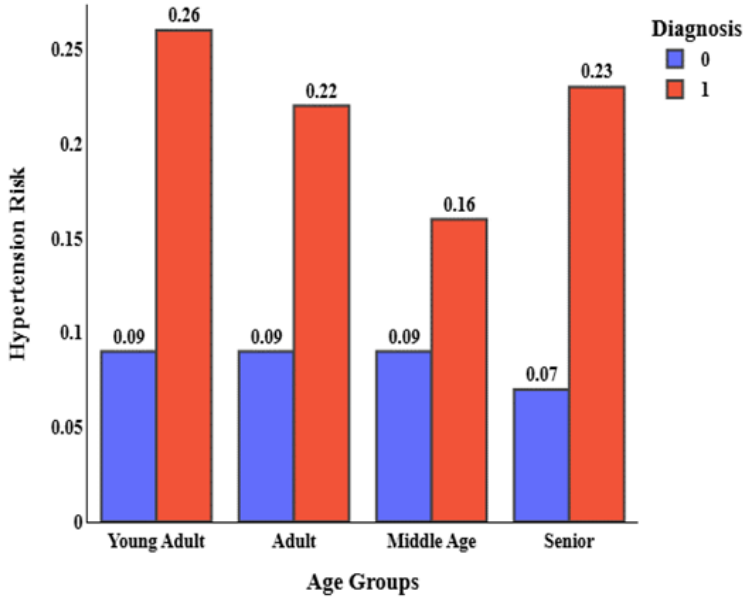

Figure 8. Hypertension Risk

Further investigation stands necessary for proving a conclusive link between liver disease and hypertension. The probability of developing liver disease depends on various factors which include individual lifestyle behaviors and alcohol consumption and genetic background. Further investigations must use longitudinal methods to track the progress of liver health in addition to changes in hypertension status throughout time.

4. Methodology

4.1 Random Forest Classifier

Random Forest classifier functions predominantly for classification together with regression work as a flexible strong ML method [34]. The method creates various decision trees during training as it calculates mean predictions for regression problems or the mode of classes for classification tasks. Bootstrap Sampling represents the initial step of Random Forest as it implements bootstrapping to produce multiple training datasets through e. sampling with substitution. Every bootstrapped dataset leads to the construction of one single decision tree. Decision trees split data points at each node by selecting random features while using a specific criterion (for classification problems Gini Impurity stands as an example). The forest combines its predictions during the conclusion phase following the

construction of every tree. The classification results from random forests select the most prevalent class among all generated trees. Automatic average computations from every tree produce the final output for regression [35].

The Gini impurity criterion determines which feature should split next in every decision tree split. The measure of Gini Impurity in node t computes as follows:

The mathematical expression comprises that pi stands for class i probability at node t and C represents the total number of groups. After splitting the primary goal is to reach the minimum level of Gini Impurity.

Random Forests employ entropy as one of its information gain criteria besides others. The entropy H(t) is given by:

$$H(t) = -\sum_{i=1}^{C} p_i \, log2(p_i) \quad (1)$$

Through the proportion of class i in node t (p_i) the equation calculates the Gini impurity criterion. The algorithm's objective aims at reducing entropy levels following every split.

4.2 CatBoostClassifier

The ML algorithm CatBoostClassifier exists specifically to process categories because they represent difficult input types for standard gradient boosting [36]. The gradient-boosting algorithm family includes CB as its member. This system was developed by Yandex under the name "Categorical Boosting." The algorithm delivers exceptional dataset generalization along with high performance levels and easy application for both numerical and categorical datasets while keeping preprocessing requirements to a minimum. The system completes automatic categorical feature preprocessing thus users can avoid manual encoding methods. Measurements called Ordered Target Statistics along with Mean Target Encoding serve as the techniques for generating separate codes for multiple categories. The encoding methodology ensures the protected confidentiality of categorical features during tree learning to stop their leak of operand information. The general gradient boosting technique faces overfitting issues with each iteration since calculating residuals introduces target value information into the model. CB implements ordered boosting to prevent data leakage during subsequent observations by first arranging data and restricting the used data to calculate residuals for each sample. Tree Construction in CB takes place one step at a time by using precedent residuals as guidance following standard boosting principles. Each decision tree operates to reduce the loss function during creation. The final prediction results from integrating all created trees. Classification tasks require ensemble trees to work together for probability prediction of all available classes. The predicted class which is likely to reveal the correct information is the last one [37].

CB builds a model through incremental training to minimize a loss function although it maintains identical gradient boosting core principles with other algorithms. Categorical data processing through systematic mechanisms and the boosting technique serve as the core things that divide CatBoost from other models.

Cross-entropy and log loss function types can be utilized as L during classification tasks.

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}) + (1 - y_i)\log(1 - \hat{y}i)] \quad (2)$$

True label y_i and prediction ŷi represent the class probability.

4.3 AdaBoost Classifier

ML features AdaBoostClassifier as its oldest and renowned boosting technique among all algorithms [38]. The abbreviation used for adaptive boosting stands as AdaBoost. The algorithm transforms inadequate learners into productive learners while improving their operational capacity. Decision stumps represent the single-split decision trees which function as weak learners during AdaBoost procedures. The model enhances total accuracy by directing its iterative weight updates toward samples which fall into ambiguous categories. AdaBoostClassifier utilizes the boosting technique that merges different weak learners which better than random guesses to create a consolidated strong learner. A process of weight adjustment leads the algorithm to earn its adaptive label. Through its adaptive process AdaBoost works on difficult classification cases before slowly adapting to specific features in the dataset. The weighting process of samples in each iteration requires that misclassified instances receive increased weight values and correctly classified instances get reduced weight values. Through this approach the algorithm devotes its computational focus to instances which received wrong classifications. The final output combines weak learners from a group of understrength student components known as the ultimate model. The weighted prediction sum represents all weak learner output predictions. Every weak learner obtains a weighting factor based on its accuracy rate. The SAMME (Stagewise Additive Modeling using a Multiclass Exponential loss function) handles multiclass classification but AdaBoostClassifier functions best when applied to binary classification problems. The basic decision trees used as weak learners in AdaBoost require minimal parameter adjustment because they are simple hence easy to manage. XGBoost and LightGBM differ from other boosting algorithms because they do not share this feature.

The weighted combination of weak learners through majority vote or summation produces the ultimate model. Binary classification predictions are

determined through the use of weighted learner prediction signs as the final prediction indicator [39].

$$F(x) = sign(\sum_{t=1}^{T} \alpha_{t.} h_t(x)) \quad (3)$$

The prediction for the T[th] weak learner is $h_t(x)$while $\alpha_{t.}$ represents its corresponding weight.

4.4 LGBM Classifier

The LGBMClassifier stands as one of the favored ML algorithms under the LightGBM (Light Gradient Boosting Machine) library which excels on massive datasets. Microsoft developed LightGBM as an open-source gradient boosting framework which uses decision trees to provide outstanding performance across regression tasks and classification tasks as well as ranking operations [40]. LGBMClassifier makes use of leaf-wise tree growth methods while performers effectively in real applications and Kaggle competitions because of its speed and memory efficiency and categorical data skills. LGBMClassifier performs different from typical tree development because it directs splitting actions to leaves which provide maximum loss reduction potential. Without max_depth regularization parameters the resulting model trees become excessively deep so overfitting occurs. The split selection process within LightGBM divides continuous features into predetermined bins through a histogram-based approach using 256 default bins for optimization. The calculation of optimal split becomes faster through this technique. LightGBM accelerates computation through its GOSS algorithm by picking random samples from low-gradient instances keeping samples from high-gradient instances. The model maintains accuracy because it directs its focus toward predicting complex examples. LightGBM merges mutually exclusive features (which usually occur separately from each other) into a single bundle known as Exclusive Feature Bundling (EFB). Diminishing the data dimensions through this approach produces faster operations without reducing information content. LGBMClassifier implements its boosting mechanism through sequential tree construction like other gradient boosting methods. A new tree in the sequential model minimizes one loss function to address errors happening in existing trees. Log loss serves as an example for classification tasks [41].

Gradient boosting serves as LGBMClassifier's foundation because it minimizes loss functions according to added sequential trees.

$$F_m(x) = F_{m-1}(x) + \gamma . h_m(x) \quad (4)$$

Gradient boosting functions according to the following general algorithm:

The model at iteration m gets represented by $F_m(x)$but the learning rate is γ and $h_m(x)$ stands for the tree produced during m that refines the earlier model predictions.

The classification loss function L(y, ŷ) amounts to log loss as its primary choice.

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}) + (1 - y_i)\log(1 - \hat{y}i)] \quad (5)$$

The true class label y_i finds its equivalent in $\hat{y}i$ as the predicted probability for class 1.

4.5 SVC

The Support Vector Classifier (SVC) serves as a strong and widely used ML algorithm for classification operations. Support Vector Machines (SVM) belong to supervised learning models which use Support Vector Classifier (SVC) to identify the most effective hyperplane that separates feature space classes [42].

The SVC algorithm identifies a hyperplane which separates the classes with the maximum possible distance during linear separation of data. The description includes a line in two dimensions, a plane in three and a hyperplane in higher-dimensional spaces. Any point specifically located next to the support vector hyperplane takes the status of support vector. The position of the hyperplane together with its margin depend on the support vectors that define them. The model structure remains unchanged regarding its relation to its support vectors when new data points are added. The kernel trick enables SVC to transpose data into elevated dimensional spaces where linear segregation becomes possible when the data cannot be split by lines. The RBF (Gaussian) kernel serves as a common choice to process non-linear data. Using the input data points as basis points the kernel function determines their product values in a higher-dimensional space without performing explicit coordinate calculation. The non-linear relationship management capabilities of SVC are enhanced because of its features. Slack variables appear in SVC to handle some classification errors which occur because the data does not divide perfectly. The misclassification tolerance levels that SVC accepts stem from the regularization parameter C. The value of C decides how smooth the decision boundary becomes because smaller C allows more misclassification points yet larger C leads to restricted misclassification tolerance.

4.6 Logistic Regression

Logistic regression functions best as a supervised learning tool in two-class problems yet allows extension to multiple classes [43]. Although its name implies regression modeling the model is in fact a classification system. It is a classification algorithm. A logistic (sigmoid) function implements probability mapping to forecast which category an observation belongs to from two available possibilities. Logistic regression operates in a linear method to resolve problems. Through suitable weighting of X={x1,x2,...,xn} features the model calculates a linear combination which

serves as an output probability prediction. The logistic regression model transforms predictive linear combinations of features through a sigmoid function instead of producing direct output values similar to linear regression methods. The sigmoid function squashes the output into a probability value between 0 and 1:

$$\sigma(z) = \frac{1}{1+e^{z}} \quad (6)$$

Where $z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$ the linear combination of the input features.

The main application of logistic regression occurs in binary classification scenarios when it calculates probabilities that lead to class predictions between 0 and 1. The predicted probability requires a threshold value of 0.5 to determine final class assignment. Two methods for extending logistic regression to multiclass classification include the One-vs-Rest (OvR) and Softmax (Multinomial Logistic Regression) approaches. The multiclass implementation of the algorithm generates class probability values before selecting the class with the

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (7)$$

Highest predicted value. Logistic regression produces mathematical models that calculate the log-odds ratio of events with class value set to 1.

A linear boundary divides the features to separate different classes in the concert space. The procedure of maximum likelihood estimation is used to establish the weights in the model. The selection process of logistic regression finds parameter values which produce the greatest possibility to observe the available data.

All machine learning models were created using Python libraries, such as scikit-learn (v1.3+), CatBoost (v1.2+), and LightGBM (v4.1+), to guarantee reproducibility and methodological transparency. To maintain the class distribution, stratified sampling was used to divide the dataset into 80% training and 20% testing sets. All models used a 4-fold stratified cross-validation approach for a thorough performance assessment.

As part of the preprocessing, categorical variables were label encoded MinMaxScaler normalization for numerical inputs used in SVC and Logistic Regression To guarantee deterministic results, a random seed of 42 was regularly used for data splitting and model initialization. The hyperparameter configurations used during model training were as follows: Random Forest (RF): n_estimators=100, max_depth=10, criterion='gini' CatBoost (CB): iterations=500, learning_rate=0.05, depth=8 AdaBoost (AB): n_estimators=50, learning_rate=1.0 LightGBM (LGBM): num_leaves=31, learning_rate=0.1, n_estimators=100 Support Vector Classifier (SVC): kernel='rbf', C=1.0, gamma='scale' Logistic Regression (LR): solver='lbfgs', C=1.0.

4.7 Methodological Flow Diagram

Figure 9 is a systematic flow of the step-by-step process that was applied here in this work of predicting liver disease. All the stages, through which raw data are collected and model is evaluated and the diseases are classified, are included in this flowchart [44]. During the Data Collection phase, the characteristics of the clinical and demographic factors that are related to the liver health is gathered with the help of a publicly available Kaggle dataset. Raw data is then followed by data preprocessing whereby the data is scrubbed, normalized and transformed to ensure consistency and adaptivity to be used in the model. Moreover, this step addresses outliers, missing values, and encoding labels when it is necessary. To ensure the objective performance estimate, the dataset is then divided into training and testing sets with the appropriate split ratios or training on cross-validation strategy. Some of the machine learning models using the evaluation & liver disease prediction is carried out by determining the performance.

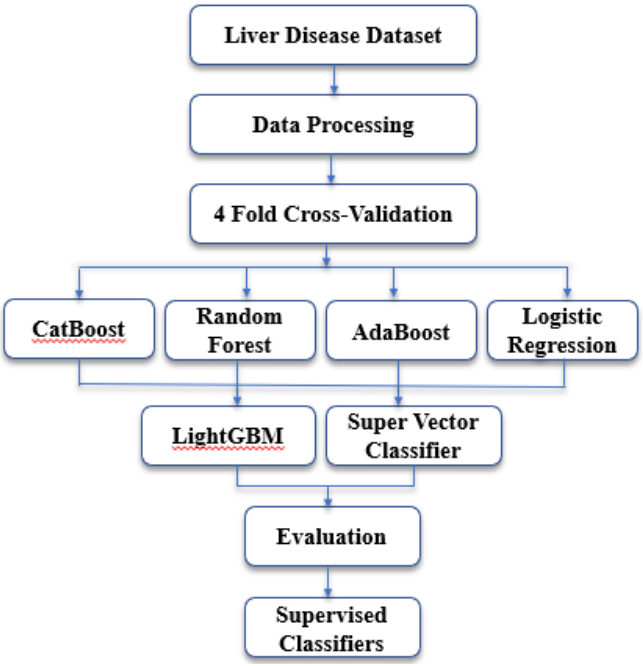

Figure 9. Step by Step Process

Even though XGBoost belongs in the number of popular and fairly effective ensemble learning algorithms, it was not used in the research due to the longer time spent in training and the overallcomplexity during the training process compared to CB and LGBM, especially with small- to medium-sized data sets. CB and LGBM were chosen as exemplars of gradient boosting models because they support faster training, better native processing of categorical features CB, more efficient memory consumption (LGBM). These characteristics made them more comfortable with the size and structure of our data.

5. Results and Discussions

This part shows the results of the analysis performed on using several machine learning models on the data of liver disease. Evaluation of each classifier is carried out on the basis of significant indicators such as ROC-AUC, F1-score, recall, accuracy, and precision. The influence and the distribution of clinical and style factors among different groups of people and the diagnostic groups are also observed with the help of data visualizations. RF, CB, ADB, LGBM, SVC and LR are compared in terms of their behavior to predict liver disease and see which of them is suitable.

5.1 Validation by Statistics

Table 2 Performance although conventional measures of performance such as F1-score, accuracy, precision and recall provide a snapshot of performance, they fail to explain how variability and uncertainty will arise when the modification of data split is employed. To address this, we performed 4- fold cross-validation to compute 95 percent confidence intervals (CI) of each measure across the fold.

The calculation of confidence intervals by assuming the t-distribution and bootstrapping allowed making a statistically valid comparison between the models. Moreover, to determine whether those differences in performance were statistically significant we performed a paired t-test on cross-validated F1-scores. Significance was determined where p-value was less than 0.05. The findings revealed that CB is much superior to other models as the difference is statistically significant ($p < 0.01$), proving its credibility in predicting the liver disease outcomes.

5.2 Comparative Evaluation

The image demonstrates how ML classifiers perform for early detection of liver disease except patient symptoms. An analysis of SV and RF together with ADB classifier, CB and LGBM classifier makes up the six models studied. The metrics used to analyze each classifier include binary classification support for positive class numbers 0 and 1 together with recall score and f1-score and precision measurements. Furthermore, metrics measure how precision (specificity) and recall (sensitivity) harmonize and their integration leads to an f1-score single metric. CB achieves the highest overall performance showing the best f1-scores as well as accuracy of 0.91 in both classes according to Table 3. With an accuracy of 0.90, RF outperforms CB, especially in class 1 (f1-score of 0.90). ADB and LGBM share identical performance levels since they achieve 0.88 accuracy with equivalent precision and recall values across every class. At 0.81 accuracy, LR trails the boosting algorithms. The lowest f1-score and accuracy rate (0.77) indicates SVC fails to produce successful results in this classification task especially when identifying cases of class 0. The better detection capabilities of CB and RF models stem from their ability to analyze intricate patterns within the data thus helping pre-symptomatic liver disease diagnosis. The medical diagnosis field benefits from these models when finding false negatives needs reduction since the models provide higher recall rates for positive class cases. Each model exhibits different sensitiveness-to-precision ratios in its performance.

The boosting techniques CB and AB deliver superior performance to all other evaluated approaches. Table 3 Classification Reports The models succeed because they can properly handle unbalanced datasets as well as performing iterative learning processes.

The ML classification model Cross Validation Scores are depicted in Figure. 10 for the investigation of pre-apparent liver disease signs. The six classification models which appear on the x-axis use these names using different machine learning algorithms. The model performance metric designated as Cross Validation Score appears on the y-axis for the evaluation. During cross-validation the average accuracy is typically displayed in such models.

Table 2. Statistical Assessment of Model

Classifier	Accuracy (Mean ± 95% CI)	F1-Score (Mean ± 95% CI)	p-value vs. CB
CatBoost	0.91 ± 0.014	0.92 ± 0.012	—
Random Forest	0.90 ± 0.018	0.90 ± 0.015	0.012
LGBM	0.88 ± 0.020	0.89 ± 0.017	0.008
AdaBoost	0.88 ± 0.021	0.89 ± 0.016	0.004
Logistic Regression	0.81 ± 0.025	0.83 ± 0.023	< 0.001
SVC	0.77 ± 0.029	0.79 ± 0.026	< 0.001

Table 3. Statistical Assessment of Model

Classifier	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Macro Avg. (F1)	Weighted Avg. (F1)
ADB	0.88	0.9	0.85	0.87	0.87	0.92	0.89	0.88	0.88
CB	0.91	0.95	0.87	0.91	0.89	0.96	0.92	0.91	0.91
LGBM	0.88	0.9	0.85	0.87	0.87	0.92	0.89	0.88	0.89
LR	0.81	0.82	0.78	0.8	0.81	0.85	0.83	0.81	0.81
RF	0.9	0.91	0.85	0.88	0.87	0.93	0.9	0.89	0.89
SVC	0.77	0.79	0.72	0.75	0.76	0.83	0.79	0.77	0.77

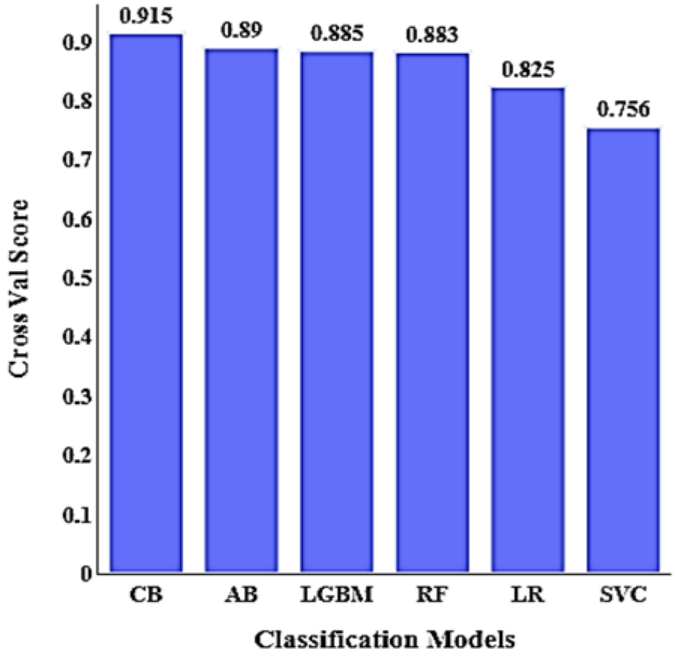

Figure 10. Cross Validation score

The cross-validation scores demonstrate that CB achieved best results across every fold while reaching the highest cross-validation score of 0.915. At 0 point the performance of AB matches CB at 0.89 while showing persistent excellent results. The accuracy measurement for the strong performer RF amounted to 0.883. Analysis revealed that LGBM achieved a cross-validation score of 0.885 even though it was slightly different from RF and LGBM. The LR model returns a score of 0.825 which demonstrates a still acceptable outcome. The SVC demonstrates the least success according to cross-validation scoring at 0.756 because of its challenged recognition accuracy throughout different validation sets. The best performing models are CB, AB, LGBM and RF because their cross-validation scores exceed 0.88.

5.3 Explaining measurement and confusion matrix

This paper includes confusion matrices of each of the classification algorithms as well as provides a detailed explanation of the performance metrics to ensure model evaluation accuracy and transparency. Once the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are shown, confusion matrix provides a graphical look at the predictions made by the classifier.

These values are used to determine the main evaluation metrics:

$$\textbf{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (8)$$

Measures how accurate the predictions performed by the model in general. The wastes that are correctly predicted to be positive as a proportion of the correctly predicted to be positive and the correct prediction of all the positivity is what is represented by the formula

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

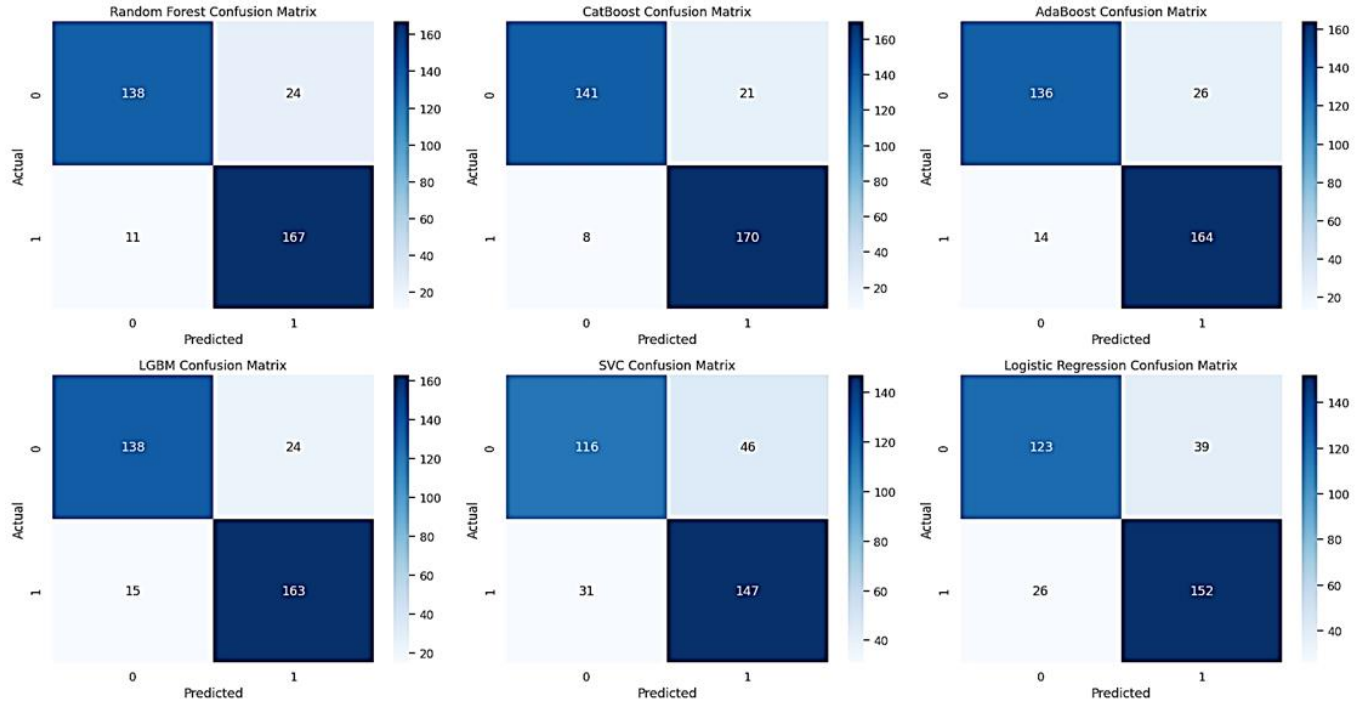

Figure 11. Confusion Matrix

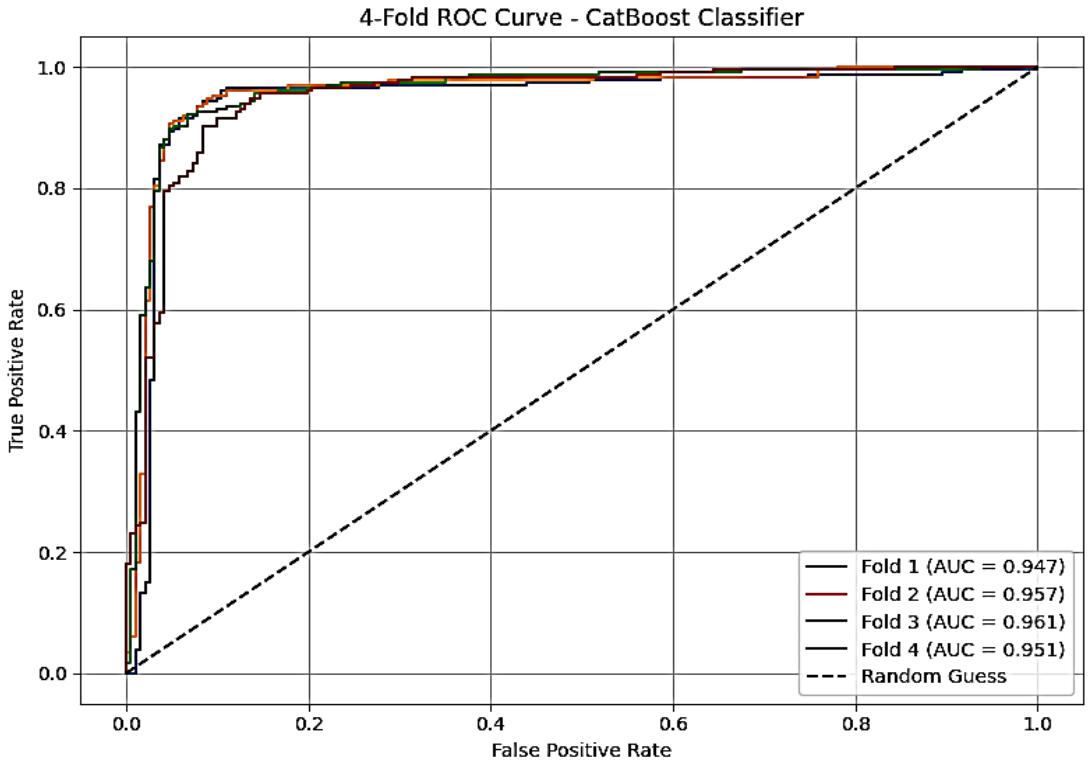

Figure 12. Four-Fold Cross Validation

The sensitivity or recall of the model can be measured by $\frac{TP}{(TP+FN)}$ that reflects an ability to detect all actual positive cases. F1-Score especially comes in handy when speaking about class and its calculation is

$$2X\frac{(\text{Precision}+\text{Recall})}{(\text{Precision x Recall})} \quad (10)$$

It is or rather, an arithmetic mean of the accuracy and recall. Table 3 presents these numbers per classifier

(RF, CB, AB, LGBM, SVC, and LR) and encompasses class-wise scores, macro and weighted averages. Moreover, Figure. 11 presents the matrices of confusion of the models in a single graphical grid, which

can be used to analyse the course of classification and the possible misclassification areas of the models. The matrices show that CatBoost and Random Forest classifiers have higher TP and TN rates, implying that they are ideal in solving the problem of identifying liver disease at an early stage.

The Figure. 12 The 4- fold cross-validation has its ROC curve in each case when the CB classifier was used to predict liver disease. The Area under the Curve (AUC) scores of Folds 1 (0.947), 2 (0.957) 3 (0.961), and 4 (0.951) were estimated separately. The CB model strength and discriminative power in identifying cases of liver diseases and non-cases are proven by the high AUC values on all folds. The diagonal line displays performance of a random classifier.

6. Conclusions

A complete research analyzes six different classification ML models through their results by evaluating SVC alongside RF, AB and CB and LR and LGBM. The cross-validation method depicts the representatives of model average performance based on multiple testing and training splits of the data. The dominant model in this study is CB since it achieves superior results than other models while maintaining the highest cross-validation score at 0.915. Under these circumstances CB emerges as the optimal choice of classifier for this particular problem. RF and LGBM demonstrate similar performance quality as CB according to cross-validation results. A discrepancy exists in terms of accuracy among the three models excluding AB, LR, and SVC. The CB system demonstrates superior performance compared to every analyzed algorithm in this investigation according to cross-validation assessments. The classification accuracy together with predicted values for new data point to CB as the superior algorithm. CB demonstrates suitable performance for this current classification project according to the obtained results. This model presents applicability across multiple cases because it identifies complex relationships while avoiding system overfitting scenarios. When selecting the ideal model for specific applications other factors including interpretation capacity and cost of computation need consideration. Future studies will benefit by applying deep learning frameworks such as transformer-based architectures, recurrent neural networks, convolutional neural networks (CNNs), although, in the present study, the focus was placed on traditional machine learning classifiers. These frameworks can have a higher accuracy and generalizability due to their usage of multimodal datasets such as time-series patient records or medical imaging as they can find arbitrary patterns in large volumes of information.

References

[1] S.K. Asrani, H. Devarbhavi, J. Eaton, P.S. Kamath, Burden of liver diseases in the world, Journal of hepatology, 70(1), (2019) 151–171.

[2] C. Castaneda, K. Nalley, C. Mannion, P. Bhattacharyya, P. Blake, A. Pecora, A. Goy, K.S. Suh, Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. Journal of Clinical Bioinformatics, 5(1), (2015) 4. https://doi.org/10.1186/s13336-015-0019-3

[3] S.M. Mahmud, M.A. Hossin, M.R. Ahmed, S.R.H. Noori, M.N.I. Sarkar, Machine Learning Based Unified Framework for Diabetes Prediction. Proceedings of the 2018 International Conference on Big Data Engineering and Technology. ACM (2018), 46-50. https://doi.org/10.1145/3297730.3297737

[4] S.N.N. Alfisahrin, T. Mantoro, (2013) Data mining techniques for optimization of liver disease classification, In 2013 International Conference on Advanced Computer Science Applications and Technologies, IEEE, Kuching, Malaysia, https://doi.org/10.1109/ACSAT.2013.81

[5] R. Moreau, M. Tonon, A. Krag, P. Angeli, M. Berenguer, A. Berzigotti, J. Fernandez, C. Francoz, T. Gustot, R. Jalan, M. Papp, EASL Clinical Practice Guidelines on acute-on-chronic liver failure. Journal of Hepatology, 79(2), (2023) 461-491. https://doi.org/10.1016/j.jhep.2024.03.012

[6] C. Hsu, C. Caussy, K. Imajo, J. Chen, S. Singh, K. Kaulback, M.D. Le, J. Hooker, X. Tu, R. Bettencourt, M. Yin, Magnetic resonance vs transient elastography analysis of patients with nonalcoholic fatty liver disease: a systematic review and pooled analysis of individual participants. Clin Gastroenterol Hepatol, Elsevier 17(4), (2019) 630–637. https://doi.org/10.1016/j.cgh.2018.05.059

[7] T. Hydes, M. Moore, B. Stuart, M. Kim, F. Su, C. Newell, D. Cable, A. Hales, N. Sheron, Can routine blood tests be modelled to detect advanced liver disease in the community: model derivation and validation using UK primary and secondary care data. British Medical Journal (BMJ), 11(2), (2021) e044952. https://doi.org/10.1136/bmjopen-2020-044952

[8] N. Rana, K. Sharma, A. Sharma, Diagnostic Strategies Using AI and ML in Cardiovascular Diseases: Challenges and Future Perspectives. In: Dulhare, U.N., Houssein, E.H. (eds) Deep Learning and Computer Vision: Models and Biomedical Applications. Algorithms for Intelligent Systems. Springer, Singapore, 1, (2025) 135-165. https://doi.org/10.1007/978-981-96-1285-7_7

[9] E.H. Houssein, M.E. Hosney, W.M. Mohamed,

A.A. Ali, E.M. Younis, Fuzzy-based hunger games search algorithm for global optimization and feature selection using medical data. Neural Computing and Application, Springer, 35, (2023) 5251–5275. https://doi.org/10.1007/s00521-022-07916-9

[10] M. Wang, S. Tang, G. Li, Z. Huang, S. Mo, K. Yang, J. Chen, B. Du, J. Xu, Z. Ding, F. Dong, Comparative study of ultrasound attenuation analysis and controlled attenuation parameter in the diagnosis and grading of liver steatosis in non-alcoholic fatty liver disease patients, BMC Gastroenterology, 24(1), (2024) 81. https://doi.org/10.1186/s12876-024-03160-8

[11] S.S. Pandi, V.R. Chiranjeevi, K.T, K.P, (2023) Improvement of Classification Accuracy in Machine Learning Algorithm by Hyper-Parameter Optimization, 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE),IEEE, Chennai, India. https://doi.org/10.1109/RMKMATE59243.2023.10369177

[12] F. Mostafa, E. Hasan, M. Williamson, H. Khan, Statistical machine learning approaches to liver disease prediction. Livers, 1(4), (2021) 294–312, https://doi.org/10.3390/livers1040023

[13] R. Amin, R. Yasmin, S. Ruhi, M.H. Rahman, M.S. Reza, Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms, Informatics in Medicine Unlocked, 36, (2023) 101155. https://doi.org/10.1016/j.imu.2022.101155

[14] M. Abdar, M. Zomorodi-Moghadam, R. Das, I.H. Ting, Performance analysis of classification algorithms on early detection of liver disease, Expert Systems. Applications, 67, (2017) 239–251. https://doi.org/10.1016/j.eswa.2016.08.065

[15] J.H. Joloudari, H. Saadatfar, A. Dehzangi, S. Shamshirband, Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection, Inform. Med. Unlocked, 17, (2019) 100255. https://doi.org/10.1016/j.imu.2019.100255

[16] S.M. Ganie, P.K. Dutta Pramanik, Predicting Chronic Liver Disease Using Boosting Technique, 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI),IEEE,Raipur, India, 1-6. https://doi.org/10.1109/ICAIIHI57871.2023.10489026

[17] Y. Wang, P. Wang, Development and validation of a new diagnostic prediction model for NAFLD based on machine learning algorithms in NHANES 2017-2020.3. Hormones (2025) 1-16. https://doi.org/10.1007/s42000-025-00634-6

[18] H. Saleem, Hepatitis Diagnosis: A Comprehensive Review of Machine Learning Classification Algorithms. The Indonesian Journal of Computer Science, 13(3), (2024).

[19] M. Wang, X. Yan, Y. Dong, X. Li, B. Gao, Machine learning and multi-omics data reveal driver gene-based molecular subtypes in hepatocellular carcinoma for precision treatment. PLoS Computational Biology, 20(5), (2024) e1012113. https://doi.org/10.1371/journal.pcbi.1012113

[20] H. Zhu, Y. Zhou, D. Shen, D.K. Wu, X. Gan, X. Xue, W. Zhang, X. Yang, J. Qiu, D. Sun An interpretable machine learning model for predicting early liver metastasis after pancreatic cancer surgery. BMC Cancer, 25, (2025) 1117. https://doi.org/10.1186/s12885-025-14503-3

[21] X. Wang, Q. Xia, S. Yang, C. Deng, N. Gu, Y. Shen, Z. Wang, B. Shi, R. Zhao, Machine Learning-Based Immuno-Inflammatory Index Integrating Clinical Characteristics for Predicting Coronary Artery Plaque Rupture. Immunity, Inflammation and Disease, 13(4), (2025) e70162. https://doi.org/10.1002/iid3.70162

[22] A. Sheakh, T. ahosin Sazia, T. aminul Islam, R.J. Lima, Improving hepatits C diagnosis using machine learning techniques: An experimental analysis. Artificial Intelligence for Intelligent Systems. CRC Press, (2024) 241-259.

[23] Q. Song, X. He, Y. Wang, H. Gao, L. Tan, J. Ma, L. Kang, P. Han, Y. Luo, K. Wang, (2025). Clinical validation of AI assisted animal ultrasound models for diagnosis of early liver trauma. Scientific Reports, 15(1), 1-9. https://doi.org/10.1038/s41598-025-91900-5

[24] Y. Wu, W. Zhao, L. Zhang, Y. Wang, Y. Wen, L. Liu, (2025). Machine learning models for predicting chemotherapy-induced adverse drug reactions in colorectal cancer patients. Digestive and Liver Disease. https://doi.org/10.1016/j.dld.2025.06.007

[25] Misc rabie_el_kharoua_2024, Predict Liver Disease: 1700 Records Dataset, https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset

[26] J. Singha, S. Baggab, R. Kaur, Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. Procedia Computer Science, Elsevier, 167, (2020) 1970–1980. https://doi.org/10.1016/j.procs.2020.03.226

[27] S. Ballestri, F. Nascimbeni, E. Baldelli, A. Marrazzo, D. Romagnoli, A. Lonardo, NAFLD as a Sexual Dimorphic Disease: Role of Gender and Reproductive Status in the Development and Progression of Nonalcoholic Fatty Liver Disease and Inherent Cardiovascular Risk. Advances in therapy, 34(6), (2017) 1291–1326. https://doi.org/10.1007/s12325-017-0556-1

[28] A. Gramenzi, F. Caputo, M. Biselli, F. Kuria, E. Loggi, P. Andreone, M. Bernardi, Review article: Alcoholic liver disease – pathophysiological aspects and risk factors. Alimentary Pharmacology & Therapeutics, 24(8), (2006) 1151-1161. https://doi.org/10.1111/j.1365-2036.2006.03110.x

[29] R.J. Shephard, N. Johnson, Effects of physical activity upon the liver. European journal of applied physiology, 115, (2015) 1–46. https://doi.org/10.1007/s00421-014-3031-6

[30] A.K. Loomis, S. Kabadi, D. Preiss, C. Hyde, V. Bonato, J. Desai, J.M. Gill, P. Welsh, D. Waterworth, N. Sattar, Body Mass Index and Risk of Nonalcoholic Fatty Liver Disease: Two Electronic Health Record Prospective Studies. The Journal of Clinical Endocrinology & Metabolism, 101(3), (2016) 945-952. https://doi.org/10.1210/jc.2015-3444

[31] S.M. Rutledge, A. Asgharpour, Smoking and Liver Disease. Gastroenterology & Hepatology, 16(12), (2020) 617. https://pmc.ncbi.nlm.nih.gov/articles/PMC8132692/

[32] J.M. Hazlehurst, C. Woods, T. Marjot, J.F. Cobbold, J.W. Tomlinson, Non-alcoholic fatty liver disease and diabetes. Metabolism, 65(8), (2016) 1096-1108. https://doi.org/10.1016/j.metabol.2016.01.001

[33] J.H. Henriksen, S. Møller, Hypertension and liver disease. Current hypertension reports, 6(6), (2004) 453–461. https://doi.org/10.1007/s11906-004-0041-5

[34] A. Paul, D.P. Mukherjee, P. Das, A. Gangopadhyay, A.R. Chintha, S. Kundu, Improved Random Forest for Classification, In IEEE Transactions on Image Processing, 27(8), (2018) 4012-4024. https://doi.org/10.1109/TIP.2018.2834830

[35] A. Cutler, D.R. Cutler, J.R. Stevens, (2012) Random Forests. Ensemble Machine Learning, Springer, New York. https://doi.org/10.1007/978-1-4419-9326-7_5

[36] A.A. Ibrahim, R.L. Ridwan, M.M. Muhammed, R.O. Abdulaziz, G.A. Saheed, Comparison of the CatBoost classifier with other machine learning methods. International Journal of Advanced Computer Science and Applications, 11(11), (2020) 738-748. https://dx.doi.org/10.14569/IJACSA.2020.0111190

[37] A.V. Dorogush, V. Ershov, A. Gulin, (2018). CatBoost: Gradient boosting with categorical features support.ArXiv. https://doi.org/10.48550/arXiv.1810.11363

[38] A. Sanusi, C.A. Putra, F.A. Akbar, Implementation of ADABOOST Algorithm on C50 for Improving the Performance of Liver Disease Classification. JEECS (Journal of Electrical Engineering and Computer Sciences), 8(2), (2023) 93-102. https://doi.org/10.54732/jeecs.v8i2.1

[39] N. Pavitha, S. Sugave, (2023). Comparative Analysis of Classification Models in Design Process of Ensemble Classifier. Information and Communication Technology for Competitive Strategies (ICTCS 2022), Lecture Notes in Networks and Systems, Springer, Singapore, 623, (2023). https://doi.org/10.1007/978-981-19-9638-2_8

[40] R. Angeline, J. Sowmiya, F. Malcom, K. Rachitha, Resume Classification Using LGBM Algorithm with Sentiment Analysis. Resume Classification Using LGBM Algorithm with Sentiment Analysis. In International Conference on Smart Computing and Communication Singapore: Springer Nature Singapore, 383-393. https://doi.org/10.1007/978-981-97-1320-2_31

[41] S.M. Ganie, P.K. Dutta Pramanik, Z. Zhao, Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches. BMC Medical Informatics and Decision Making, 24(1), (2024)160. https://doi.org/10.1186/s12911-024-02550-y

[42] C.C. Chang, C.J. Lin, Training v-Support Vector Classifiers: Theory and Algorithms, In Neural Computation, 13(9), (2001) 2119-2147. https://doi.org/10.1162/089976601750399335

[43] I. Naglik, M. Lango, (2025) Fine-Tuning Fine-Tuned Models: Towards a Practical Methodology for Sentiment Analysis with Small In-Domain Supervised Dataset. In International Conference on Neural Information Processing, Singapore, Springer Nature Singapore, 1-16. https://doi.org/10.1007/978-981-96-7005-5_1

[44] H. Vu-Ngoc, S.S. Elawady, G.M. Mehyar, A.H. Abdelhamid, O.M. Mattar, O. Halhouli, N.L. Vuong, C.D. Mohd Ali, U.H. Hassan, N.D. Kien, K. Hirayama, N.T. Huy, Quality of flow diagram in systematic review and/or meta-analysis. PLOS ONE, 13(6), (2018) e0195955. https://doi.org/10.1371/journal.pone.0195955

Authors Contribution Statement

Sumathi Selveraj: Conceptualization, methodology, formal analysis, writing original draft. Ramesh Thenappan: Data collection, investigation, Writing, review and editing. Parthiban Chakkaravarthi: Data analysis, Data curation, validation, supervision, Writing, review and editing. All the authors read and approved the final version of this work.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License