CLASSIFYING LIVER DISEASE WITH BOOSTING MACHINE LEARNING APPROACHES

Ümit YILMAZ[1*], Erol ÖZÇEKİÇ[2]

[1] Balıkesir Üniversitesi, Bigadiç Meslek Yüksekokulu, Yönetim ve Organizasyon Bölümü, Balıkesir
[2] Balıkesir Üniversitesi, Balıkesir Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Balıkesir
[1] ORCID No: https://orcid.org/0000-0003-4268-8598
[2] ORCID No: https://orcid.org/0000-0002-1896-6853

Keywords
Liver disease
Machine learning
Diagnosis
Classification
Boosting

Abstract

Liver diseases pose a significant global health challenge due to their impact on metabolic function and the difficulty of early detection. Traditional diagnostic methods such as liver biopsy have limitations due to their invasive nature and high costs. This research examines the application of advanced machine learning techniques such as Gradient Boosting, AdaBoost, XGBoost and CatBoost for classification of liver diseases using a publicly available dataset of 1700 clinical records. Statistical analyses identified key predictors such as age, body mass index (BMI), lifestyle factors, and liver function tests, which were used to train and evaluate the models. The performance of the models was evaluated using metrics such as accuracy, precision, recall and AUC-ROC. The CatBoost model showed the highest performance with an accuracy of 93.82%, while also producing the most consistent results with precision (91.97%), recall (96.62%), F1 score (94.25%) and AUC-ROC (95.64%). These results highlight the potential of machine learning-based approaches to improve diagnostic accuracy and reduce reliance on invasive procedures. The proposed framework can contribute to improving patient outcomes and optimizing healthcare resources by providing a foundation for real-time clinical decision support systems.

BOOSTING MAKİNE ÖĞRENME YAKLAŞIMLARI İLE KARACİĞER HASTALIKLARININ SINIFLANDIRILMASI

Anahtar Kelimeler
Karaciğer hastalığı
Makine öğrenmesi
Tanı
Sınıflandırma
Boosting

Öz

Karaciğer hastalıkları, metabolik fonksiyonlar üzerindeki etkileri ve erken teşhis zorlukları nedeniyle önemli bir küresel sağlık sorunu oluşturmaktadır. Karaciğer biyopsisi gibi geleneksel tanı yöntemleri, invaziv yapıları ve yüksek maliyetleri nedeniyle sınırlamalar taşımaktadır. Bu araştırma, 1.700 klinik kayıttan oluşan kamuya açık bir veri kümesi kullanarak karaciğer hastalıklarının sınıflandırılması için Gradient Boosting, AdaBoost, XGBoost ve CatBoost gibi gelişmiş makine öğrenmesi tekniklerinin uygulanmasını incelemektedir. İstatistiksel analizler, modelleri eğitmek ve değerlendirmek için kullanılan yaş, vücut kitle endeksi, yaşam tarzı faktörleri ve karaciğer fonksiyon testleri gibi temel belirleyicileri ortaya koymuştur. Modellerin performansı, doğruluk, kesinlik, duyarlılık ve AUC-ROC gibi metrikler kullanılarak değerlendirilmiştir. CatBoost modeli, %93,82 doğruluk oranı ile en yüksek performansı göstermiş, aynı zamanda kesinlik (%91,97), duyarlılık (%96,62), F1 skoru (%94,25) ve AUC-ROC (%95,64) değerleriyle en istikrarlı sonuçları üretmiştir. Bu sonuçlar, makine öğrenmesi tabanlı yaklaşımların tanı doğruluğunu artırma ve invaziv prosedürlere olan bağımlılığı azaltma potansiyelini vurgulamaktadır. Önerilen çerçeve, gerçek zamanlı klinik karar destek sistemleri için bir temel oluşturarak hasta sonuçlarının iyileştirilmesine ve sağlık hizmetleri kaynaklarının optimizasyonuna katkı sağlayabilir.

Araştırma Makalesi — Research Article
Başvuru Tarihi : 27.11.2024 — Submission Date : 27.11.2024
Kabul Tarihi : 25.04.2025 — Accepted Date : 25.04.2025

* Corresponding author: umityilmaz@balikesir.edu.tr
https://doi.org/10.31796/ogummf.1591951

1. Introduction

The liver performs numerous vital functions essential to human metabolism and the maintenance of homeostasis. These include the regulation of glucose levels, lipid metabolism, protein synthesis, detoxification of harmful substances, and the production of bile necessary for digestion (Casotti & D'Antiga, 2019). Liver diseases encompass a wide spectrum of conditions that impair these functions, posing significant global health challenges. Early diagnosis and accurate classification of liver diseases are essential for effective treatment and management, as delayed diagnosis can lead to severe complications, including cirrhosis and liver failure (Decharatanachart et al., 2021).

Traditional diagnostic methods often rely on invasive procedures and subjective interpretations, which may require significant time and are prone to inaccuracies. Despite the availability of various noninvasive tests, liver biopsy remains the gold standard in clinical practice. However, biopsy is a highly invasive and risky procedure, associated with significant costs and patient discomfort. Moreover, inter- and intra-observer variability cannot be avoided, even when biopsy is performed by experienced pathologists. The associated morbidity and mortality rates strongly suggest that the use of this procedure should be reduced (Decharatanachart et al., 2021).

In recent years, the emergence of ML has created new opportunities for enhancing diagnostic accuracy and efficiency in medical practice. This research explores the application of various ML algorithms, as Gradient Boosting, AdaBoost, XGBoost, and CatBoost, in the classification of liver diseases. By leveraging these advanced computational techniques, the research endeavors to improve predictive performance and provide a strong framework for the early identification of liver-related ailments. Through a comparative analysis of these algorithms, the research endeavors to identify the most effective approach for classifying liver disease, ultimately contributing to better patient results and more informed clinical decision-making (Jovović et al., 2024).

The integration of ML in liver disease diagnosis holds promise for reducing reliance on invasive procedures and minimizing diagnostic errors. By analyzing complex datasets and identifying patterns not easily discernible through traditional methods, these algorithms can facilitate earlier and more accurate detection of liver diseases. This approach not only enhances patient care but also optimizes resource utilization within healthcare systems (Yadav & Singhal, 2023).

Ultimately, the use of ML techniques in the classification and diagnosis of liver diseases points out a significant advancement in medical science. By improving diagnostic accuracy and enabling early intervention, these technologies have the potential to transform liver disease management and improve patient results (Ganie et al., 2024).

2. Literature Review

Liver diseases encompass a broad spectrum of conditions that significantly impact global health. Conventional diagnostic approach, such as liver biopsy, are invasive and carry associated risks, underscoring the need for non-invasive, accurate diagnostic tools. In recent years, ML approaches have emerged as promising alternatives for the classification and diagnosis of liver diseases. This literature review examines various ML techniques applied to liver disease classification, highlighting their methodologies, datasets, and performance outcomes.

Mathur et al. (2024) conducted a study on classifying fatty liver disease (FLD) using various ML approaches, including XGBoost, Naive Bayes, Random Forest (RF), and Artificial Neural Networks (ANNs). These algorithms were trained on well-curated datasets to predict the incidence of FLD. The performance of the models was evaluated using accuracy scores and confusion matrices, showcasing their effectiveness in accurately predicting FLD. Notably, a hybrid model combining ANN and XGBoost achieved the highest accuracy of 90.25%, outperforming the other individual algorithms tested. The study highlights the potential of ML, particularly hybrid models, in facilitating early detection and improved management of FLD, which is crucial for mitigating disease progression.

Pilankar and Shyamala (2024) conducted a study on classifying liver diseases using ensemble learning techniques, integrating multiple supervised ML models, including ANNs, Decision Tree (DT), Gradient Boosting, Logistic Regression (LR), RF and Support Vector Machine (SVM). By employing extensive hyperparameter tuning and feature selection, the research achieved a notable accuracy rate of 79%.

Hidayat (2024) conducted a study utilizing the K-Nearest Neighbors (KNN) algorithm to classify liver diseases based on biochemical markers derived from a dataset of 584 patient records. The dataset included key variables such as age, gender, and results from various liver function tests. The model's performance was assessed using cross-validation, yielding an accuracy range of 57.76% to 73.28%.

Raj et al. (2024) conducted a study aimed at classifying liver diseases using multiple ML algorithms, including RF, LR, DT, SVM, and KNN. The dataset used for the study consisted of 583 instances and 11 features, with a gender distribution of 142 females and 441 males. Among the algorithms tested, the RF classifier

demonstrated the best performance, achieving an accuracy of 72%.

Aouragh and Bahaj (2023) conducted a comparative study evaluating multiple ML algorithms (SVM, RF, Extra Trees (ET), Gradient Boosting) for classifying liver diseases. The research emphasized the importance of preprocessing, feature selection, and dimensionality reduction techniques in optimizing model performance. Using an unbalanced dataset comprising 583 patient records, the developed models demonstrated promising results, achieving accuracy exceeding 91.60% with SVM.

Makkena and Natarajan (2023) investigated the application of various ML algorithms for liver disease classification, focusing on models such as RF, KNN, XGBoost, DT, LR, SVM, and ET classifiers. To address class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, enhancing the algorithms' performance. Among the evaluated models, XGBoost demonstrated the highest accuracy at 72%, followed by ET at 70.2%, and RF at 67.4%. The study highlights the effectiveness of XGBoost in handling imbalanced datasets for liver disease classification while emphasizing the importance of advanced preprocessing techniques like SMOTE in improving ML outcomes.

Mohamed et al. (2023) conducted a study on liver disease classification using various ML algorithms. The research utilized a Kaggle dataset comprising 441 male and 142 female patient records, with ten features to distinguish between cases of liver disease (1) and non-liver disease (0). The study assessed the performance of multiple ML models, including LR, SVM, Gradient Boosting, DT, and RF. Among these, the RF algorithm achieved the highest accuracy at 80.34%, followed by LR (76.07%), SVM (74.36%), Gradient Boosting (71.79%), and DT (70.09%).

The reviewed studies demonstrate the growing application and effectiveness of ML techniques in classifying and diagnosing liver diseases. From hybrid models like ANN-XGBoost achieving remarkable accuracy to the effective handling of imbalanced datasets using methods like SMOTE, these approaches illustrate the versatility and adaptability of ML in addressing complex diagnostic challenges. Ensemble learning methods, preprocessing techniques, and feature selection have consistently emerged as critical factors in optimizing model performance. While traditional algorithms like RF and SVM continue to show reliable results, advancements in hybrid and ensemble methodologies highlight the potential for even greater diagnostic accuracy. Collectively, these studies highlight the revolutionary potential of ML in liver disease classification, paving the way for more precise, non-invasive, and efficient diagnostic tools in clinical practice.

3. Materials and Methods

3.1. Dataset

This study used a dataset to demonstrate the use of ML in health data research. In this study, the "Predict Liver Disease: 1700 Records Dataset" was utilized. This dataset is open access and freely available for researchers on the Kaggle platform (El Kharoua, 2024). As the study was conducted using a publicly available clinical dataset, obtaining Ethics Committee approval was not required. The dataset comprises 1,700 individuals, with 10 predictors and one response variable (liver disease (LD) diagnosis: No/Yes), as detailed in Table 1. However, it is not known whether this label represents a general liver disease or a specific type. The dataset was complete, containing no missing values. Each data record in the dataset is described with 10 features that could have predictive power, including age, gender, BMI, smoking, alcohol consumption score, physical activity score, genetic risk, hypertension, diabetes, liver function test. The gender distribution among participants was nearly equal, with 843 males (49.59%) and 857 females (50.41%). The mean age of participants was 50.39±17.64 years, providing a representative distribution for the study of liver disease risk factors and diagnostic features.

The dataset was split into training and test sets using an 80/20 ratio, resulting in 1,360 training samples and 340 test samples. This split was chosen based on standard machine learning practices, ensuring sufficient data for model training while maintaining a reliable test set for evaluation.

The initial basic statistical analyses involved applying the Shapiro-Wilk normality test to assess the distribution of quantitative predictors in the dataset ($p<0.05$). Based on the type of predictors, the Mann-Whitney U test was used for non-parametric quantitative variables, and the Pearson chi-square test was employed for categorical variables to identify statistically significant differences between LD diagnosis groups. A significance level of $p \leq 0.05$ was set as the threshold for type I error. All statistical analyses were performed using IBM SPSS Statistics software (version 20.0.0).

Table 1. Summary of Descriptive and Inferential Statistics for Predictor Variables

	Predictor*		LD diagnosis No (n=764)	LD diagnosis Yes (n=936)	p**
Demographics	Age		45 (30)	54 (29)	<0.001
	Gender	Male	459 (60.08%)	384 (41.03%)	<0.001
		Female	305 (39.92%)	552 (58.97%)	
Lifestyle Indicators	BMI		25.65 (13)	29.19 (11)	<0.001
	Smoking	No	618 (80.89%)	586 (62.61%)	
		Yes	146 (19.11)	350 (37.39%)	
	Alcohol Consumption Score		6.48 (7)	12.53 (9)	<0.001
	Physical Activity Score		5.36 (5)	4.62 (5)	<0.001
Patient History	Genetic Risk	Low	459 (60.08%)	519 (55.45%)	<0.001
		Medium	278 (36.39%)	279 (29.81%)	
		High	27 (3.53%)	138 (14.74%)	
	Hypertension	No	698 (91.36%)	739 (78.95%)	<0.001
		Yes	66 (8.64%)	197 (21.05%)	
	Diabetes	No	687 (89.92)	771 (82.37%)	<0.001
		Yes	77 (10.08%)	165 (17.63%)	
Laboratory Measurements	Liver Function Test		45.38 (35)	68.47 (32)	<0.001

* The distribution of quantitative variables was presented using median values and interquartile ranges, whereas categorical variables were summarized by the number and proportion of individuals.

** The Mann-Whitney U and Pearson chi-square tests were used to calculate p-values for numerical and categorical predictors, respectively.

3.2. Machine Learning Algorithms

Artificial intelligence (AI) can create a meaningful impact on the early prediction and detection of diseases by enabling regular, effective, and critical analysis of medical data (Kumar et al., 2023). ML algorithms have been instrumental in developing models that analyze multimodal data, facilitating intelligent and cost-effective patient monitoring, early intervention, effective treatment planning, and timely management (Pei et al., 2023). In the healthcare context, ML processes predominantly involve supervised and unsupervised learning (An et al., 2023). ML methodologies are predominantly utilized for applications like classification, regression, and clustering analysis (Sarker, 2021).

Unlike traditional statistical modeling, ML-based, model-free algorithms can react to changes in new data and can be enabled to take intelligent actions to learn or develop an adaptive model that is capable of carrying out specific tasks. This is particularly beneficial for dealing with large and complex medical data. Furthermore, ML technology can be employed to forecast the need for clinical and other support services and can offer insight into the effectiveness of therapy, the input factors for a therapeutic regimen, and hospital or patient readmission rates by using data from medical records (Bennett et al., 2022). Healthcare organizations are increasingly leveraging ML technologies to address variability in healthcare utilization and spending patterns. By employing regression analysis, these organizations can identify areas for cost savings and optimize resource allocation. Additionally, ML aids in determining appropriate sample sizes for cost-effective study designs, ensuring robust and efficient research outcomes. Collectively, these advanced capabilities contribute to substantial cost reductions in the healthcare system while enhancing the quality and delivery of care (Anderson et al., 2022).

In this study, four ML algorithms are employed: Gradient Boosting, AdaBoost, XGBoost, and CatBoost. These algorithms are widely recognized for their effectiveness in various predictive modeling tasks. To ensure a fair comparison, all models were trained and tested using the same training and test split. No additional preprocessing was applied separately for different models, and the same feature set was used across all algorithms.

All ML analyses were performed using Python software (version 3.13.0), leveraging its extensive libraries and tools for data preprocessing, model training, evaluation, and visualization.

To optimize the performance of the machine learning models, hyperparameter tuning was performed using Grid Search Cross Validation (GridSearchCV). A 5-fold cross-validation approach was employed to systematically explore the best hyperparameter combinations for each model. The optimal parameters were selected based on the highest accuracy achieved on the validation set. Table 2 summarizes the final hyperparameters used for each model.

Table 2. Optimized Hyperparameters for Models

Model	Learning Rate	Max Depth	Number of Estimators
GradientBoosting	0.1	3	50
AdaBoost	0.5	-	50
XGBoost	0.1	3	100
CatBoost	0.1	10	100

3.2.1. Gradient Boosting

Gradient Boosting is a powerful ML technique that constructs predictive models by sequentially aggregating several base learners, commonly decision trees, to create a robust ensemble model. This method

sequentially adds subsequent models to correct errors made by existing ensemble, thereby improving overall performance (Ayyadevara, 2018).

The algorithm operates by minimizing a specified loss function through gradient descent in function space. At each iteration, it trains each new model to approximate the negative gradient of the loss function, calculated based on the current ensemble's predictions. This approach allows the model to focus on areas where previous models performed poorly, effectively reducing bias and variance (Biau et al., 2019).

Mathematically, the Gradient Boosting algorithm can be described as follows:

Initialization: Start with a constant model that minimizes the loss function:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \quad (1)$$

where $L(y_i, \gamma)$ is the loss function, y_i are the true values, and n is the number of training samples.

Iterative Process: For $m = 1$ to M (the total number of iterations):

Compute Residuals: Calculate the pseudo-residuals (negative gradients) for each sample:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad (2)$$

Fit Base Learner: Train a weak learner $h_m(x)$to predict the residuals r_{im}.

Compute Multiplier: Determine the optimal multiplier γ_m by solving:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \quad (3)$$

Update Model: Update the ensemble model:

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

Final Model: After M iterations, the final model is:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x) \quad (5)$$

This formulation highlights the additive nature of Gradient Boosting, where each new model incrementally improves the performance of the ensemble by focusing on the errors of its predecessors (Bentéjac et al., 2021). Gradient Boosting has been proficiently utilized in diverse fields, like classification, regression, and ranking tasks, because of its flexibility and high predictive accuracy (Luo et al., 2023).

3.2.2. AdaBoost

AdaBoost (Adaptive Boosting) is a prominent ensemble learning algorithm that aggregates several base classifiers to create a robust classifier with improved predictive performance. Introduced by Freund and Schapire in 1995, AdaBoost iteratively adjusts the weights of training samples, emphasizing those that were misclassified in previous iterations, thereby directing subsequent weak learners to focus on the more challenging cases (Schapire, 2013).

The AdaBoost algorithm operates as follows:

Initialization: Assign equal weights to all training samples.

Iterative Process (for $t = 1$ to T, where T is the total number of iterations):

Train a weak classifier $h_t(x)$ using the current weights.

Compute the weighted error ϵ_t of $h_t(x)$:

$$\epsilon_t = \frac{\sum_{i=1}^{N} w_i \cdot I(y_i \neq h_t(x_i))}{\sum_{i=1}^{N} w_i} \quad (6)$$

where w_i is the weight of sample i, y_i is the true label, x_i is the feature vector, and $I(\cdot)$ is the indicator function.

Calculate the classifier's weight (α_t):

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (7)$$

Update the weights of the training samples:

$$w_i \leftarrow w_i \cdot \exp\left(-\alpha_t \cdot y_i \cdot h_t(x_i)\right) \quad (8)$$

Normalize the weights so that they sum to one.

Final Hypothesis: Combine the weak classifiers into a final strong classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right) \quad (9)$$

AdaBoost has been widely applied in various domains, including object detection and recognition. For instance, Viola and Jones utilized AdaBoost to develop a rapid object detection framework, achieving real-time face detection with high accuracy (Schapire, 2013). Additionally, AdaBoost has been employed in text classification tasks, demonstrating its versatility across different data types (Bozuyla, 2021).

Despite its strengths, AdaBoost can be sensitive to noisy data and outliers, as it tends to assign higher weights to misclassified samples, possibly resulting in overfitting. To tackle this, several modifications and extensions have been proposed, such as incorporating regularization techniques and developing robust versions of the algorithm (Hornyák & Iantovics, 2023). In summary, AdaBoost is a foundational ensemble learning method that enhances the performance of weak classifiers through adaptive weighting and iterative training, making it a valuable tool in the ML practitioner's toolkit.

3.2.3. XGBoost

XGBoost, which stands for eXtreme Gradient Boosting, is a highly optimized and scalable variant of the traditional gradient boosting algorithm, designed to improve both

the speed and performance of machine learning models. It has gained significant attention for its effectiveness and accuracy in diverse predictive applications tasks (Bentéjac et al., 2021).

XGBoost operates by constructing decision trees sequentially, with each new tree trained to reduce the residual errors generated by the ensemble up to that point. This iterative process minimizes a specified loss function, leading to a robust predictive model.

Mathematically, the objective function in XGBoost is defined as:

$$L(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \quad (10)$$

where n is number of training instances, l is differentiable convex loss function that evaluates the gap of the observed y_i relative to the predicted $\hat{y}_i^{(t)}$ values, t is iteration step, f_k is the k-th tree in the ensemble and $\Omega(f_k)$ is regularization term for the k-th tree.

The regularization term $\Omega(f_k)$ is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (11)$$

where T represents leaves' number in the tree, w_j represents leaf's weight j, γ and λ represent regularization parameters. This regularization helps control the complexity of the model, preventing overfitting and enhancing generalization (Bentéjac et al., 2021).

XGBoost incorporates several innovative features to improve its performance. It efficiently handles sparse data through a sparsity-aware algorithm, making it particularly beneficial for datasets with missing values or zero entries. Additionally, it employs an approximate tree learning algorithm known as the weighted quantile sketch, which enables the handling of weighted data and facilitates the construction of trees with weighted quantiles. XGBoost also leverages parallel processing and distributed computing techniques, accelerating training and ensuring it highly ideal for high-volume datasets. These advancements collectively demonstrate that XGBoost is a all-round and robust approach, capable of delivering high performance throughout a varied range of scenarios (Chen & Guestrin, 2016). These enhancements make XGBoost a versatile and powerful tool in the ML practitioner's arsenal, capable of delivering high performance across a wide range of applications.

3.2.4. CatBoost

CatBoost is a gradient boosting framework introduced by Yandex, known for its ability to handle categorical features natively and its mechanisms that help prevent overfitting. It introduces innovative techniques such as ordered boosting and efficient processing of categorical variables, enhancing model accuracy and robustness.

CatBoost incorporates several innovative techniques that distinguish it from traditional gradient boosting methods. One notable feature is its ability to address prediction shift caused by target leakage, a common issue where information from the target variable inadvertently influences the model during training. To counteract this, CatBoost implements ordered boosting, a sequential data processing approach that ensures each data point is used for training only after its target value has been predicted. This method effectively prevents target leakage and reduces the risk of overfitting.

Another significant advancement in CatBoost is its efficient handling of categorical variables. Unlike traditional methods that require extensive preprocessing steps such as one-hot encoding, CatBoost introduces a novel algorithm known as "ordered target statistics." This technique replaces categorical values with statistics derived from the target variable, calculated in a manner that avoids target leakage. By leveraging this approach, CatBoost can efficiently manage high-cardinality categorical features, enhancing its performance and usability in datasets with complex categorical variables (Prokhorenkova et al., 2018).

CatBoost builds an ensemble of decision trees in a stage-wise manner, aiming to minimize a specified loss function. The model at the t-th iteration, $F_t(x)$, is updated as follows:

$$F_t(x) = F_{t-1}(x) + \gamma_t h_t(x) \quad (12)$$

where $F_{t-1}(x)$ is the ensemble model from the previous iteration, $h_t(x)$ is is the newly added decision tree, trained to approximate the negative gradient of the loss function at iteration t, γ_t is the learning rate, controlling the contribution of $h_t(x)$ to the ensemble.

The negative gradient, g_{ti}, for the i-th instance at iteration t is computed as:

$$g_{ti} = -\left[\frac{\partial L\left(y_i, F_{t-1}(x_i)\right)}{\partial F_{t-1}(x_i)}\right] \quad (13)$$

where $L\left(y_i, F_{t-1}(x_i)\right)$ is the loss function evaluated at the true target y_i and the model's prediction $F_{t-1}(x_i)$. The new decision tree $h_t(x)$ is then trained to predict g_{ti} for each instance i.

CatBoost has demonstrated superior performance across various domains, including finance, healthcare, and e-commerce, particularly in tasks involving datasets with numerous categorical features. Comparative studies have shown that CatBoost often outperforms other gradient boosting implementations, such as XGBoost and LightGBM, in terms of accuracy and generalization capabilities (Bentéjac et al., 2021). In summary, CatBoost's innovative handling of categorical features and its approach to mitigating prediction shift

make it a powerful tool for ML practitioners dealing with complex datasets.

3.3. Feature Selection

"There are many features that researchers use for detecting liver disease, and here in this research, 10 features are considered. Feature selection is a critical step in building supervised classifiers. A wrongly selected feature can lead a ML algorithm to develop a biased model, where for some features, the weight multiplier may be compromised. In addition, various studies show that a high number of factors usually cause inaccurate and unstable performance.

Feature selection for liver classifiers can help by maintaining the relevance, quality, and interpretability of the model. Improving classification performance decreases training and execution time and also increases clarity. Therefore, for selecting classification features, a method that can evaluate various features differently is desirable. In this study, XGBoost algorithm is used for feature selection. After feature selection, all features were found to be effective in predicting liver disease. The result of feature selection is shown in Figure 1.

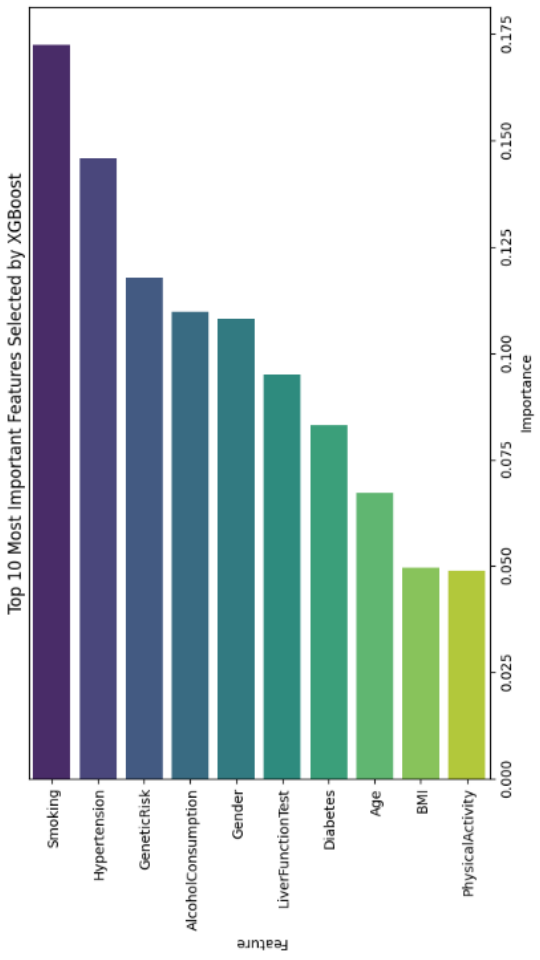

Figure 1. The Selected Predictors by the XGBoost Feature Importance Feature Selection Algorithm

The feature importance analysis conducted using XGBoost revealed that smoking, hypertension, genetic risk, and alcohol consumption were the most influential features in liver disease classification. These variables demonstrated a strong predictive impact, significantly contributing to model performance. Additionally, gender, liver function tests, diabetes, and age were identified as moderately important predictors. In contrast, BMI and physical activity had relatively lower importance in the model's decision-making process. These findings suggest that certain lifestyle and genetic factors play a more dominant role in liver disease prediction. The insights derived from feature selection provide a clearer understanding of the underlying patterns in the dataset, enhancing model interpretability and supporting the development of data-driven clinical decision-making frameworks."

3.4. Evaluation Matrix

To evaluate the performance of the ML algorithms, several evaluation metrics such as accuracy, precision, recall, F1 score, and the AUC-ROC are used in this study.

The dataset was split into training and test sets using an 80/20 ratio, resulting in 1,360 training samples and 340 test samples. However, to ensure a fair comparison, all machine learning models were trained and tested on the same dataset partition.

Additionally, cross-validation techniques, such as 5-fold cross-validation with GridSearchCV, were used for hyperparameter tuning. However, the final evaluation of model performance was conducted on the same fixed test set to ensure consistency in model comparison.

Accuracy represents the ratio of correctly predicted instances to the total instances in a dataset and is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

where TP, TN, FP, and FN represent the numbers of true positive, true negative, false positive, and false negative outcomes, respectively (Ghosh et al., 2021).

Precision, recall, and F1 score are important metrics in classification tasks used to evaluate class-wise performance.

Precision is the fraction of relevant instances among the retrieved instances and is defined as follows:

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

where TP and FP denote the number of correctly predicted true positives and falsely predicted false positives, respectively (Hinojosa Lee et al., 2024).

Recall is the fraction of the total amount of relevant instances found over the total amount of instances and is defined as:

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

where TP and FN denote the number of correctly predicted true positives and falsely predicted false negatives, respectively.

The F1 score, also known as the harmonic mean of precision and recall, is calculated using the following formula (Rewicki et al., 2023):

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (17)$$

The AUC-ROC measures the ability of the model to distinguish between classes. AUC is area under the curve plotted with the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis (Verbakel et al., 2020). These are calculated as:

$$TPR\ (Recall) = \frac{TP}{TP+FN} \quad (18)$$

$$FPR = \frac{FP}{FP+TN} \quad (19)$$

In this study, research and publication ethics were followed.

4. Results

Table 1 summarizes the descriptive statistics of the predictor variables stratified by liver disease status, along with their corresponding p-values. The results indicate that all variables exhibit statistically significant differences between the diagnostic groups.

Table 2 presents the confusion matrix derived from the Gradient Boosting model evaluated on the test dataset. Out of 340 test samples, the model accurately predicted 306 cases and misclassified 34 instances.

Table 2. Confusion Matrix of the Gradient Boosting Model on the Test Dataset

Model Prediction (LD)	Actual Diagnosis (LD)		Total
	No	Yes	
No	**140**	*12*	152
Yes	*22*	**166**	188
Total	162	178	340

Note: Cells containing bolded values indicate correct classifications, whereas italicized values represent misclassifications within the confusion matrices.

Table 3 displays the confusion matrix for the AdaBoost model evaluated on the test set. Among 340 test instances, the model correctly classified 308 samples and incorrectly predicted 32 cases.

Table 3. Confusion Matrix of the AdaBoost Model on the Test Dataset

Model Prediction (LD)	Actual Diagnosis (LD)		Total
	No	Yes	
No	**142**	*12*	154
Yes	*20*	**166**	186
Total	162	178	340

Note: Cells containing bolded values indicate correct classifications, whereas italicized values represent misclassifications within the confusion matrices.

Table 4 presents the confusion matrix of the XGBoost model applied to the test data. Out of 340 observations, the model accurately classified 309 instances, while 31 were incorrectly predicted.

Table 4. Confusion Matrix of the XGBoost Model on the Test Dataset

Model Prediction (LD)	Actual Diagnosis (LD)		Total
	No	Yes	
No	**141**	*10*	151
Yes	*21*	**168**	189
Total	162	178	340

Note: Cells containing bolded values indicate correct classifications, whereas italicized values represent misclassifications within the confusion matrices.

Table 5 displays the confusion matrix for the CatBoost model evaluated on the test dataset. Of the 340 test samples, the model correctly identified 319 cases and produced 21 misclassifications.

Table 5. Table 4. Confusion Matrix of the CatBoost Model on the Test Dataset

Model Prediction (LD)	Actual Diagnosis (LD)		Total
	No	Yes	
No	**147**	*6*	153
Yes	*15*	**172**	187
Total	162	178	340

Note: Cells containing bolded values indicate correct classifications, whereas italicized values represent misclassifications within the confusion matrices.

Table 6 provides a comparative assessment of LD classification performance across the Gradient Boosting, AdaBoost, XGBoost, and CatBoost models, based on

several evaluation metrics. The performance scores were derived from the confusion matrices shown in Tables 2 to 5.

Table 6. LD Classification Results of the Models

Metric	GradientBoosting	AdaBoost	XGBoost	CatBoost
Accuracy	0.900000	0.905882	0.908824	**0.938235**
Precision	0.882979	0.892473	0.888889	**0.919786**
Recall	0.932584	0.932584	0.943820	**0.966292**
F1 Score	0.907104	0.912088	0.915531	**0.942466**
ROC-AUC	0.950149	0.948276	0.955091	**0.956374**

Note: Cells with bold numbers best results

Figure 2 illustrates the comparative AUC-ROC performance of the Gradient Boosting, AdaBoost, XGBoost, and CatBoost models in the classification of LD.

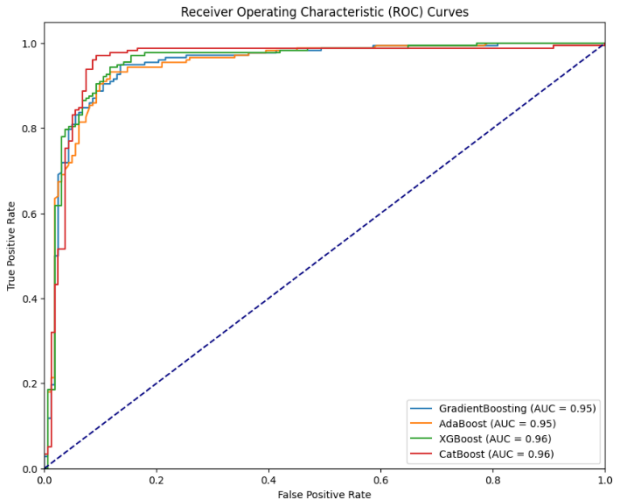

Figure 2. AUC-ROC Results of the Models

5. Conclusions

Liver disease is a significant public health issue. Identifying whether a patient has liver disease is challenging and often requires multiple diagnostic tests, including blood tests, imaging (such as ultrasound, MRI, or CT scans), and, in some cases, invasive liver biopsy. The shortage of well-qualified specialists in hepatology can contribute to delays in diagnosis and treatment, which, if left unaddressed, may lead to disease progression and severe complications such as cirrhosis or liver failure.

To enhance the accuracy of liver disease classification and support clinical decision-making, this study developed and evaluated machine learning (ML) models using clinical data. The dataset utilized in this study consisted of 1,700 patient records with 10 key features, split into training and testing sets in an 80/20 ratio. Four ML models—Gradient Boosting, AdaBoost, XGBoost, and CatBoost—were trained and tested using performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. Among the models, CatBoost achieved the highest accuracy (93.82%), while Gradient Boosting had the lowest accuracy (90.00%). The findings suggest that ML-based models can effectively classify liver disease, with performance differences across models being relatively minor.

The findings of this study hold significant implications for practical applications in healthcare. The proposed ML models can be integrated into clinical decision support systems to assist physicians in diagnosing liver disease more efficiently and accurately. While these models do not replace traditional diagnostic methods such as biopsy, they can serve as complementary tools to improve early detection and reduce unnecessary invasive procedures. Furthermore, ML-based diagnostic tools could be deployed in telemedicine platforms, enabling early screening and proactive disease management, particularly in regions with limited access to hepatology specialists. By incorporating this approach into routine clinical workflows, healthcare professionals could benefit from faster, more data-driven decision-making, ultimately improving patient outcomes. However, to ensure real-world applicability, future studies should focus on validating these models using primary clinical datasets and evaluating their performance in hospital settings.

One of the key limitations of this study is the use of a secondary dataset obtained from Kaggle. While this dataset provides a valuable foundation for ML model development, it may have undergone preprocessing or modifications that are not explicitly documented, potentially introducing biases. Additionally, publicly available datasets may not fully capture the complexities of real-world clinical scenarios. Future research should validate the proposed approach using hospital-acquired datasets and assess its generalizability in real clinical environments.

Additionally, an 80/20 train-test split was applied. While this approach is widely used, it may introduce minor variations in class distributions between training and test sets. Furthermore, all models were trained and tested on the same predefined dataset split. Although this ensures a fair comparison, alternative validation techniques such as k-fold cross-validation or holdout validation could further enhance the robustness and generalizability of the results. In future studies, these techniques should be used to strengthen model evaluation.

This study opens several avenues for future research to enhance liver disease diagnosis. While this study focused on evaluating the performance of boosting-based machine learning approaches in liver disease

classification, future research could explore comparisons with conventional machine learning models such as RF, SVM, and LR. This could provide a broader perspective on the relative effectiveness of boosting algorithms in the context of liver disease diagnosis. Future efforts could focus on employing data augmentation techniques to improve early-stage detection capabilities and integrating multimodal data, such as imaging and genomic markers, to refine model accuracy. Additionally, developing explainable AI models would provide interpretable insights for clinicians, fostering trust and facilitating clinical adoption. Testing the framework on larger, diverse, and multiethnic datasets would ensure its generalizability and fairness. Finally, longitudinal studies incorporating patient data over time and the real-time deployment of these models in clinical settings could significantly advance the integration of ML in liver disease management.

Author Contributions

Ümit YILMAZ and Erol ÖZÇEKİÇ contributed to the publication with the design and implementation of the research, analysis of the results, and writing of the manuscript and discussion of the results and review of the manuscript.

Conflict of Interest

There is no conflict of interest with any person/institution in the prepared article

References

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors*, *23*(9), 4178. https://www.mdpi.com/1424-8220/23/9/4178

Anderson, D., Bjarnadottir, M. V., & Nenova, Z. (2022). Machine Learning in Healthcare: Operational and Financial Impact. In V. Babich, J. R. Birge, & G. Hilary (Eds.), *Innovative Technology at the Interface of Finance and Operations: Volume I* (pp. 153-174). Springer International Publishing. https://doi.org/10.1007/978-3-030-75729-8_5

Aouragh, A. A., & Bahaj, M. (2023, 16-22 Dec. 2023). Feature Selection and Dimensionality Reduction for Unbalanced Liver Disease Classification with Optimized Machine Learning Algorithms. 2023 7th IEEE Congress on Information Science and Technology (CiSt),

Ayyadevara, V. K. (2018). Gradient Boosting Machine. In *Pro Machine Learning Algorithms : A Hands-On Approach to Implementing Algorithms in Python and R* (pp. 117-134). Apress. https://doi.org/10.1007/978-1-4842-3564-5_6

Bennett, M., Hayes, K., Kleczyk, E. J., & Mehta, R. (2022). Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics. *arXiv preprint arXiv:2201.02469*.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

Biau, G., Cadre, B., & Rouvière, L. (2019). Accelerated gradient boosting. *Machine Learning*, *108*(6), 971-992. https://doi.org/10.1007/s10994-019-05787-1

Bozuyla, M. (2021). AdaBoost Ensemble Learning on top of Naive Bayes Algorithm to Discriminate Fake and Genuine News from Social Media [Naive Bayes Algoritmasının AdaBoost Topluluk Öğrenme Modeli ile Sosyal Medyada Sahte ve Gerçek Haberlerinin Ayırt Edilmesi]. *European Journal of Science and Technology*(28), 459-462. https://doi.org/10.31590/ejosat.1005577

Casotti, V., & D'Antiga, L. (2019). Basic Principles of Liver Physiology. In L. D'Antiga (Ed.), *Pediatric Hepatology and Liver Transplantation* (pp. 21-39). Springer International Publishing. https://doi.org/10.1007/978-3-319-96400-3_2

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Decharatanachart, P., Chaiteerakij, R., Tiyarattanachai, T., & Treeprasertsuk, S. (2021). Application of artificial intelligence in chronic liver diseases: a systematic review and meta-analysis. *BMC Gastroenterology*, *21*(1), 10. https://doi.org/10.1186/s12876-020-01585-5

El Kharoua, R. (2024). *Predict Liver Disease: 1700 Records Dataset*. https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset

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

Ghosh, S., Chatterjee, A., & Chatterjee, D. (2021). An Improved Load Feature Extraction Technique for Smart Homes Using Fuzzy-Based NILM. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-9. https://doi.org/10.1109/TIM.2021.3095093

Hidayat, A. (2024). Predictive Modelling of Liver Disease Using Biochemical Markers and K-Nearest Neighbors Algorithm. *International Journal of Artificial Intelligence in Medical Issues*, *2*(2), 104-114.

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Applied Sciences*, *14*(21), 9863. https://www.mdpi.com/2076-3417/14/21/9863

Hornyák, O., & Iantovics, L. B. (2023). AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics. *Mathematics*, *11*(8), 1801. https://www.mdpi.com/2227-7390/11/8/1801

Jovović, I., Grebović, M., Poković, L. G., Popović, T., & Čakić, S. (2024). Liver Diseases Classification Using Machine Learning Algorithms. In A. Badnjević & L. Gurbeta Pokvić, *MEDICON'23 and CMBEBIH'23* Cham.

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Humaniz Comput*, *14*(7), 8459-8486. https://doi.org/10.1007/s12652-021-03612-z

Luo, J., Wei, Z., Man, J., & Xu, S. (2023). TRBoost: a generic gradient boosting machine based on trust-region method. *Applied Intelligence*, *53*(22), 27876-27891. https://doi.org/10.1007/s10489-023-05000-w

Makkena, K. R., & Natarajan, K. (2023). Classification Algorithms for Liver Epidemic Identification. *EAI Endorsed Transactions on Pervasive Health and Technology*, *9*. https://doi.org/10.4108/eetpht.9.4379

Mathur, S., Karodi, P., & Dhanare, R. (2024, 13-14 March 2024). Fatty Liver Disease Prediction Through Machine Learning. 2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL),

Mohamed, S., Ezzat, R., Ghorab, S., Bhatnagar, R., & Shams, M. Y. (2023, 1-3 Nov. 2023). Liver Disease Identification Based on Machine Learning Algorithms. 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS),

Pei, X., Zuo, K., Li, Y., & Pang, Z. (2023). A Review of the Application of Multi-modal Deep Learning in Medicine: Bibliometrics and Future Directions. *International Journal of Computational Intelligence Systems*, *16*(1), 44. https://doi.org/10.1007/s44196-023-00225-6

Pilankar, A., & Shyamala, L. (2024, 5-7 June 2024). Liver disease Prediction using Ensemble Learning. 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC),

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Raj, H., N, G., Kodipalli, A., & Rao, T. (2024, 28-29 June 2024). Prediction of Chronic Liver Disease Using Machine Learning Algorithms and Interpretation with SHAP Kernels. 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS),

Rewicki, F., Denzler, J., & Niebling, J. (2023). Is It Worth It? Comparing Six Deep and Classical Methods for Unsupervised Anomaly Detection in Time Series. *Applied Sciences*, *13*(3), 1778. https://www.mdpi.com/2076-3417/13/3/1778

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Schapire, R. E. (2013). Explaining AdaBoost. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (pp. 37-52). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5

Verbakel, J. Y., Steyerberg, E. W., Uno, H., De Cock, B., Wynants, L., Collins, G. S., & Van Calster, B. (2020). ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*, *126*, 207-216. https://doi.org/10.1016/j.jclinepi.2020.01.028

Yadav, H. S., & Singhal, R. K. (2023, 3-5 March 2023). Classification and Prediction of Liver Disease Diagnosis Using Machine Learning Algorithms. 2023 2nd International Conference for Innovation in Technology (INOCON),