International Journal of Artificial Intelligence in Medical Issues

Volume 4 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v4i1.435

Research Article

Gender-Aware Prediction of Liver Disease Using Machine Learning and Clinical Laboratory Data

Umar Zaky [1,*]; Muhammad Habibi [2]; Adri Priadana [3]; Thomas Edyson Tarigan [4]
[1] Universitas Teknologi Yogyakarta, Daerah Istimewa Yogyakarta 55285, Indonesia, umarzaky@uty.ac.id
[2] Universitas Jenderal Achmad Yani Yogyakara, Daerah Istimewa Yogyakarta 55293, Indonesia, muhammadhabibi27@gmail.com
[3] Universitas Jenderal Achmad Yani Yogyakara, Daerah Istimewa Yogyakarta 55293, Indonesia, adripriadana3202@gmail.com
[4] Universitas Teknologi Digital Indonesia, Kota Denpasar, Bali 80225, Indonesia, tarigan@utdi.ac.id

Correspondence should be addressed to Umar Zaky; umarzaky@uty.ac.id

Received 20 January 2026; Revised 06 February 2026; Accepted 25 May 2026; Published 30 May 2026

Abstract:
Liver disease is a major health problem that may progress silently and lead to severe clinical complications if not detected early. Machine learning offers a promising approach for supporting early screening by identifying predictive patterns from clinical and biochemical patient data. This study developed an explainable gender-aware machine learning framework for liver disease prediction using demographic information and clinical biomarkers. The dataset consisted of 570 patient records after duplicate removal, including age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, SGPT, SGOT, total protein, albumin, albumin/globulin ratio, and liver disease status. Several machine learning algorithms were evaluated under three experimental scenarios: original data, class-weighted learning, and SMOTENC-based oversampling. Model performance was assessed using accuracy, precision, recall, specificity, F1-score, and ROC-AUC. The experimental results showed that Gradient Boosting combined with SMOTENC achieved the best F1-score, with an accuracy of 0.7632, precision of 0.7935, recall of 0.9012, specificity of 0.4242, F1-score of 0.8439, and ROC-AUC of 0.7759. The model correctly identified 73 of 81 liver disease cases in the testing set, indicating strong sensitivity for early screening. Gender-based evaluation showed comparable F1-scores for male and female patients, with values of 0.8430 and 0.8462, respectively. Feature importance analysis identified SGOT, alkaline phosphatase, age, and direct bilirubin as the most influential predictors. These findings suggest that an explainable and gender-aware machine learning approach can support liver disease risk prediction using routinely available clinical biomarkers, although further validation using larger and more balanced datasets is required.

Keywords: Liver Disease Prediction, Machine Learning, Clinical Biomarkers, SMOTENC, Gender-Aware Evaluation, Explainable AI.

Dataset link: -

1. Introduction

Liver disease remains a major global health concern because it can progress silently from mild hepatic dysfunction to chronic liver disease, cirrhosis, liver failure, and other severe complications [1]. The burden of liver cirrhosis and other chronic liver diseases has increased substantially over the last decades, with global deaths rising from approximately 1.01 million in 1990 to 1.47 million in 2019. This condition is associated with multiple etiologies, including hepatitis B, hepatitis C, alcohol consumption, non-alcoholic fatty liver disease, and other metabolic or inflammatory causes [2], . Therefore, early identification of liver disease risk is essential to support timely intervention and reduce the possibility of disease progression [3].

In clinical practice, liver function tests and related biochemical markers play an important role in detecting and monitoring liver disorders [4]. Commonly used indicators include bilirubin, alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, albumin, and total protein. These markers can reflect different patterns of liver injury, such as hepatocellular damage, cholestatic injury, impaired protein synthesis, or abnormalities in bilirubin metabolism. However, interpretation of these markers may be complex because abnormal values do not always directly indicate a specific liver disease, and clinical interpretation often requires consideration of multiple laboratory results together with patient characteristics.

Machine learning has become a promising approach for supporting clinical decision-making because it can identify complex patterns from multidimensional medical data [5]. In liver disease prediction, machine learning models can learn relationships between demographic factors, biochemical markers, and disease status to assist early screening and risk classification [6], [7]. Compared with manual interpretation alone, predictive models may provide additional support by integrating several clinical variables simultaneously [8]. However, in medical classification tasks, model development should not only focus on overall accuracy, but also consider sensitivity, specificity, class imbalance, and subgroup-level performance to ensure that the model is reliable for practical clinical use.

The dataset used in this study consists of patient records from the North-East region of Andhra Pradesh, India, and includes demographic and biochemical features such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, SGPT, SGOT, total protein, albumin, and albumin/globulin ratio. The classification task is to predict whether a patient has liver disease based on these clinical indicators [9], [10]. The dataset contains 583 patient records, with 416 patients diagnosed with liver disease and 167 patients without liver disease, indicating an imbalanced class distribution. In addition, the dataset includes 441 male and 142 female patients, making gender-based performance analysis relevant for evaluating whether the predictive model behaves consistently across patient subgroups.

Although machine learning has been widely applied to liver disease prediction, several important issues remain. First, imbalanced class distribution may cause the model to favor the majority class, resulting in poor detection of patients from the minority class [11]. Second, a model that performs well overall may still show unequal performance across gender groups. This issue is important in medical artificial intelligence because a clinically useful model should be evaluated not only by aggregate metrics, but also by its consistency across relevant patient sub-populations. Third, predictive performance alone is insufficient in healthcare applications, since clinicians and researchers need to understand which clinical features contribute most strongly to model predictions [12].

Therefore, this study proposes an explainable and gender-aware machine learning framework for liver disease prediction using clinical biomarkers [13], [14]. Several machine learning algorithms are compared to identify the most effective predictive model. Class imbalance is addressed using appropriate resampling or weighting strategies, and model performance is evaluated using accuracy, precision, recall, F1-score, specificity, and ROC-AUC. Furthermore, gender-based evaluation is conducted to examine whether the model demonstrates consistent predictive performance between male and female patients. Finally, feature importance analysis is used to improve model interpretability and identify the most influential clinical biomarkers associated with liver disease prediction.

The main contributions of this study are threefold. First, it develops and compares machine learning models for liver disease prediction using demographic and biochemical clinical features. Second, it evaluates the impact of class imbalance handling on predictive performance. Third, it incorporates gender-based performance analysis and model explainability to provide a more clinically relevant evaluation framework for AI-assisted liver disease screening [15].

2. Method

Research Design

This study employed a quantitative experimental approach to develop and evaluate machine learning models for liver disease prediction. The research was designed as a supervised binary classification task, in which clinical and biochemical patient data were used to predict whether a patient had liver disease or no liver disease. The methodological framework consisted of several stages, including dataset preparation, data preprocessing, exploratory data analysis, model development, class imbalance handling, model evaluation, gender-based performance analysis, and model interpretability analysis [16], [17].

The general workflow of this study is presented in Figure 1. The process began with collecting and preparing the liver patient dataset, followed by data cleaning and feature transformation. The cleaned dataset was then divided into training and testing subsets. Several machine learning algorithms were trained and evaluated under different experimental scenarios, including the original dataset, class-weighted learning, and oversampling using SMOTENC. Finally, the best-performing model was further analyzed based on gender-specific performance and feature importance.

Data Description:

The dataset used in this study consisted of medical records of liver patients from the North-East region of Andhra Pradesh, India. The dataset contained demographic information and biochemical laboratory test results that are commonly associated with liver function [18], [19]. The initial dataset consisted of 583 patient records and 11 variables. The target variable indicated whether a patient was diagnosed with liver disease or not.

The independent variables included age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, total protein, albumin, and albumin/globulin ratio. The dependent variable was the liver disease status.

The variables used in this study are summarized in **Table 1**.

Table 1. Description of Dataset Variables

Variable	Description	Type
Age	Age of the patient	Numerical
Gender	Gender of the patient	Categorical
TB	Total bilirubin level	Numerical
DB	Direct bilirubin level	Numerical

Variable	Description	Type
Alkphos	Alkaline phosphatase level	Numerical
Sgpt	Alanine aminotransferase level	Numerical
Sgot	Aspartate aminotransferase level	Numerical
TP	Total protein level	Numerical
ALB	Albumin level	Numerical
A/G Ratio	Albumin/globulin ratio	Numerical
Selector	Target class indicating liver disease status	Categorical

The target variable consisted of two classes: Liver Disease and No Liver Disease. Before preprocessing, the dataset contained 416 liver disease cases and 167 non-liver disease cases. This distribution indicated that the dataset was imbalanced, with the liver disease class representing the majority of the samples [20].

Data Preprocessing:

Data preprocessing was performed to improve data quality and prepare the dataset for machine learning model development. First, the dataset was examined to identify missing values, duplicate records, inconsistent labels, and data type issues. The dataset did not contain missing values. However, 13 duplicate records were identified and removed to avoid repeated samples influencing the learning process. After duplicate removal, 570 patient records remained for model development.

The target variable was converted into binary form. The Liver Disease class was encoded as `1`, while the No Liver Disease class was encoded as `0`. The gender variable was also transformed into numerical representation, where female patients were encoded as `0` and male patients were encoded as `1`.

Numerical features were standardized using standard scaling. Standardization was applied to ensure that all numerical variables had comparable scales, especially because several algorithms such as support vector machine, logistic regression, and k-nearest neighbors are sensitive to feature magnitude. The transformation followed the standard score formula:

$$z = \frac{x - \mu}{\sigma}$$

where x represents the original feature value, μ represents the mean value of the feature, and σ represents the standard deviation.

Exploratory Data Analysis:

Exploratory data analysis was conducted to understand the general characteristics of the dataset. The analysis included examining the distribution of the target classes, gender distribution, target class distribution by gender, and

correlation among numerical clinical features. This step was important to identify class imbalance, potential subgroup differences, and relationships among liver function biomarkers [21].

The class distribution analysis showed that the number of patients with liver disease was higher than the number of patients without liver disease. This imbalance could affect the predictive model by causing it to become biased toward the majority class. Therefore, class imbalance handling was included as part of the experimental design.

Data Splitting:

The dataset was divided into training and testing sets using an 80:20 split ratio. The training set was used for model development and cross-validation, while the testing set was used for final model evaluation. Stratified splitting was applied to preserve the proportion of liver disease and no liver disease cases in both subsets.

In addition, a 5-fold stratified cross-validation strategy was applied to the training data. Stratified cross-validation was used to ensure that each fold maintained a similar class distribution, thereby producing a more reliable estimate of model performance.

Machine Learning Models:

Several machine learning algorithms were compared in this study to identify the most effective model for liver disease prediction. The selected algorithms included both linear and non-linear classifiers. The models evaluated were:

- Logistic Regression, used as a baseline linear classifier because of its interpretability and common use in medical prediction tasks.
- Decision Tree, used because it can model non-linear relationships and provide interpretable decision rules.
- Random Forest, used as an ensemble learning method that combines multiple decision trees to improve prediction stability and reduce overfitting.
- Support Vector Machine, used because of its ability to construct optimal decision boundaries for classification tasks, especially in small to medium-sized datasets.
- K-Nearest Neighbors, used as a distance-based classifier for comparative evaluation.
- Gradient Boosting, used as an ensemble model that builds sequential weak learners to improve predictive performance.

These models were selected to provide a comparative evaluation of different machine learning approaches for clinical tabular data.

Class Imbalance Handling:

Because the dataset showed an imbalanced class distribution, this study evaluated three experimental scenarios.

The first scenario used the original dataset without any imbalance handling. This scenario served as the baseline condition.

The second scenario applied class-weighted learning, where higher weight was assigned to the minority class during model training. This strategy allowed the classifier to give more attention to the underrepresented class.

The third scenario applied Synthetic Minority Oversampling Technique for Nominal and Continuous Features, or SMOTENC. SMOTENC was selected because the dataset consisted of both numerical and categorical features. This method generated synthetic samples for the minority class while preserving the categorical structure of the gender feature.

By comparing these three scenarios, this study evaluated whether imbalance handling improved the model's ability to detect liver disease and no liver disease cases more fairly.

Model Evaluation:

Model performance was evaluated using several classification metrics, namely accuracy, precision, recall or sensitivity, specificity, F1-score, and ROC-AUC [22]. These metrics were selected because accuracy alone is insufficient for imbalanced medical datasets. In clinical prediction tasks, sensitivity is especially important because it measures the model's ability to correctly identify patients with liver disease. Specificity is also important because it measures the model's ability to correctly identify patients without liver disease.

The evaluation metrics were calculated as follows:

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

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

$$Recall / Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

The final model comparison was performed using the testing set. The best-performing model was selected primarily based on F1-score, with ROC-AUC used as an additional supporting metric. F1-score was prioritized because it balances precision and recall, making it suitable for imbalanced classification problems.

Gender-Based Performance Analysis:

To evaluate whether the selected model performed consistently across patient subgroups, gender-based performance analysis was conducted. The testing data were divided into male and female patient groups, and the best-performing model was evaluated separately for each group.

For each gender group, accuracy, precision, recall, specificity, F1-score, and ROC-AUC were calculated. This analysis aimed to identify whether the model showed performance differences between male and female patients. Such subgroup evaluation is important in medical artificial intelligence because a model with good overall performance may still produce unequal performance across demographic groups.

Model Interpretability Analysis:

To improve the transparency of the predictive model, feature importance analysis was conducted using permutation importance. This method evaluates the contribution of each feature by measuring the decrease in model performance after the values of a feature are randomly shuffled. A larger decrease in performance indicates that the feature has a stronger contribution to the model's prediction.

Permutation importance was selected because it can be applied to different machine learning models and does not depend on the internal structure of a specific algorithm. The analysis was used to identify which clinical biomarkers had the greatest influence on liver disease prediction [23].

The interpretability analysis is important because medical prediction models should not only produce accurate results, but also provide insight into the clinical variables that contribute to the prediction. In this study, feature importance analysis was expected to support a more transparent understanding of how demographic and biochemical variables influenced liver disease classification.

Experimental Environment:

All experiments were implemented using Python programming language. The main libraries used in this study included pandas and NumPy for data processing, scikit-learn for model development and evaluation, imbalanced-learn for class imbalance handling, and Matplotlib for data visualization. The experimental pipeline was designed to ensure reproducibility through a fixed random state in data splitting, cross-validation, resampling, and model training.

3. Result and Discussion

Dataset Characteristics:

After data preprocessing, the dataset consisted of 570 patient records and 11 variables. The original dataset contained 583 records, but 13 duplicate records were identified and removed to avoid repeated samples influencing the learning process. No missing values were found in the dataset after cleaning.

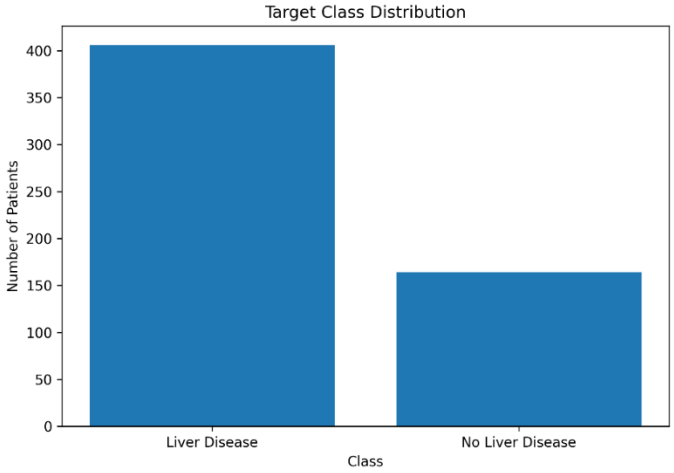

Figure 1. Target Class Distribution

The target class distribution is shown in **Figure 1**. The dataset contained 406 patients with liver disease, representing 71.23% of the total data, and 164 patients without liver disease, representing 28.77%. This distribution indicates a clear class imbalance, where the liver disease class was the majority class. In medical prediction tasks, such imbalance can affect model performance because a classifier may learn to prioritize the majority class and perform poorly in identifying the minority class.

Table 2. Target Class Distribution

Class	Count	Percentage
Liver Disease	406	71.23%
No Liver Disease	164	28.77%

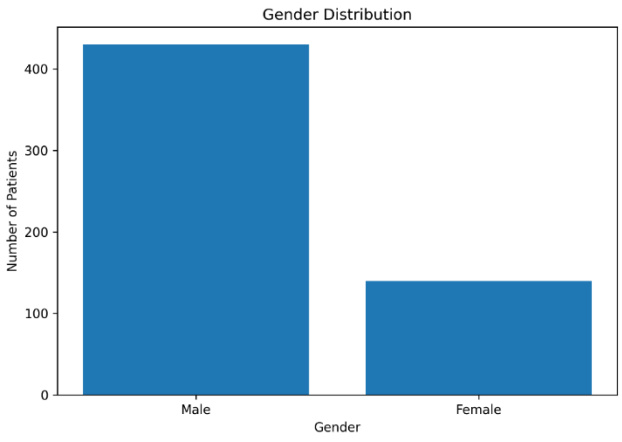

Figure 2. Gender Distribution

As shown in **Figure 2**, the dataset was also imbalanced in terms of gender. There were 430 male patients, representing 75.44% of the dataset, and 140 female patients, representing 24.56%. This unequal distribution shows that male patients were more dominant in the dataset. Therefore, gender-based performance analysis was important to evaluate whether the model produced consistent predictions across male and female patients.

Table 3. Gender Distribution

Gender	Count	Percentage
Male	430	75.44%
Female	140	24.56%

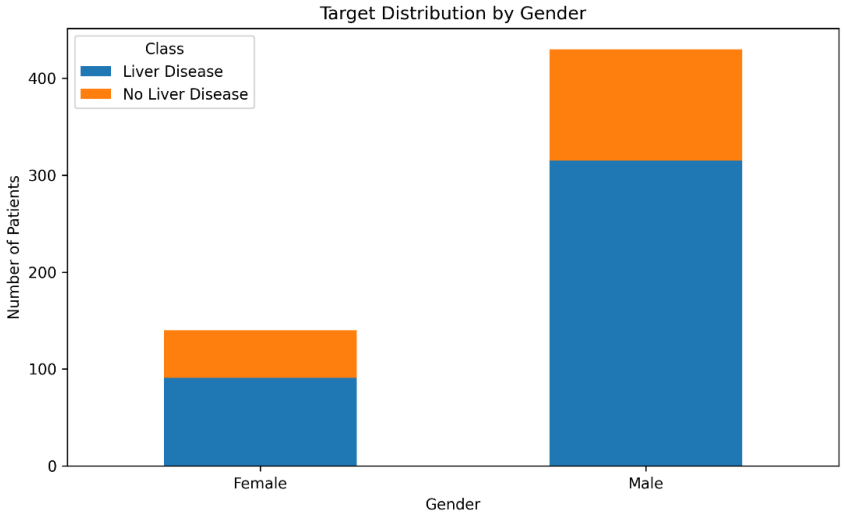

Figure 3. Target Distribution by Gender

The distribution of liver disease cases by gender is presented in **Figure 3**. Among female patients, 91 cases were classified as liver disease and 49 cases as no liver disease. Among male patients, 315 cases were classified as liver disease and 115 cases as no liver disease. In percentage terms, 65.00% of female patients and 73.26% of male patients were recorded as having liver disease. This result suggests that liver disease cases were proportionally higher among male patients in this dataset.

Table 4. Target Class Distribution by Gender

Gender	Liver Disease	No Liver Disease
Female	91	49
Male	315	115

These findings confirm two important characteristics of the dataset. First, the target variable was imbalanced because liver disease cases were more frequent than non-liver disease cases. Second, the gender distribution was also imbalanced, with male patients being substantially more represented than female patients. These conditions justify the use of imbalance handling and gender-based evaluation in this study.

Correlation Analysis of Clinical Features:

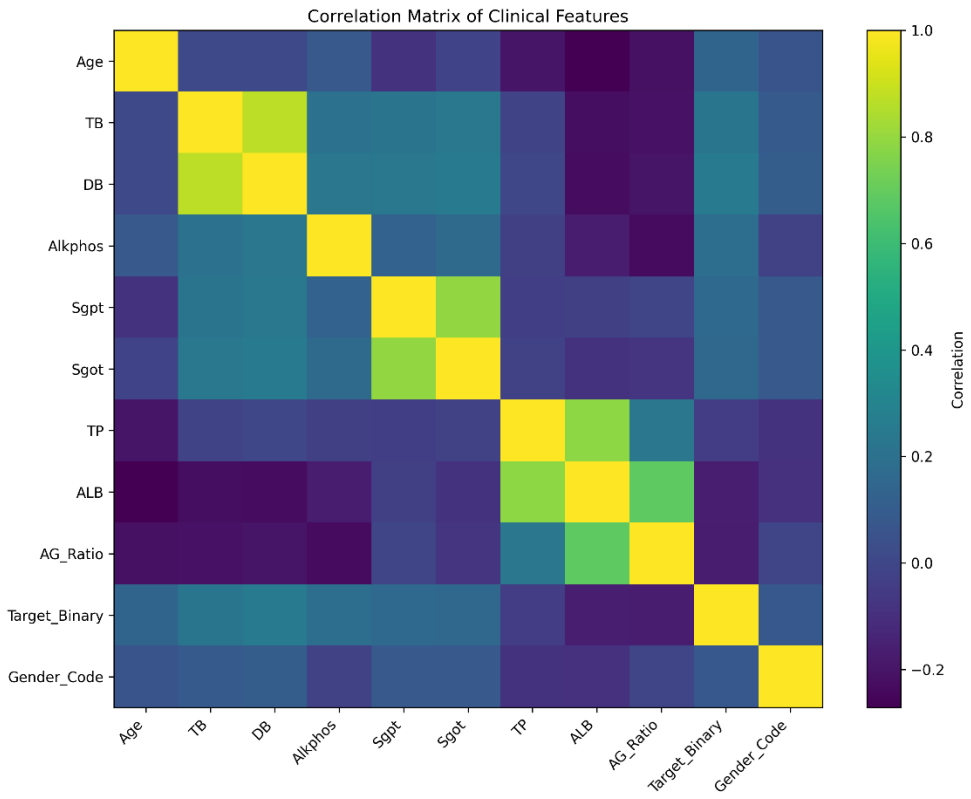

Figure 4. Correlation Matrix of Clinical Features

The correlation matrix of the clinical variables is shown in **Figure 4**. Several variables demonstrated strong relationships with each other. Total bilirubin and direct bilirubin showed a strong positive correlation, which is clinically reasonable because direct bilirubin is a component of total bilirubin. SGPT and SGOT also showed a relatively strong positive correlation, indicating that both enzyme markers may reflect related patterns of liver cell injury.

Total protein, albumin, and albumin/globulin ratio also showed visible correlations. Albumin and albumin/globulin ratio were positively associated, which is expected because albumin contributes directly to the albumin/globulin ratio. In contrast, some liver disease-related indicators showed weak or negative relationships with protein-related features. These correlation patterns indicate that the dataset contains both overlapping and complementary clinical information.

The correlation between individual clinical features and the target variable was generally moderate to weak. This suggests that liver disease prediction cannot rely on a single laboratory marker alone. Instead, machine learning models are useful because they can combine multiple demographic and biochemical variables to identify predictive patterns that may not be obvious from individual feature relationships.

Model Performance Comparison:

The dataset was divided into training and testing subsets using an 80:20 stratified split. The training set consisted of 456 records, while the testing set consisted of 114 records. The class distribution was preserved in both subsets. The training set contained 325 liver disease cases and 131 no liver disease cases, while the testing set contained 81 liver disease cases and 33 no liver disease cases.

Several machine learning models were evaluated under three experimental scenarios: original data, class-weighted learning, and SMOTENC oversampling. The models were compared using accuracy, precision, recall, specificity, F1-score, and ROC-AUC. The overall testing results are summarized in **Table 5**.

Table 5. Testing Performance of Machine Learning Models

Scenario	Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC
SMOTENC	Gradient Boosting	7,632	7,935	9,012	4,242	8,439	7,759
Original	SVM	7,105	7,105	1.2000	0	8,308	6,902
Original	Gradient Boosting	7,193	7,333	9,506	1,515	8,280	7,561
ClassWeight	Random Forest	7,193	7,379	9,383	1,818	8,261	7,720
Original	Random Forest	7,193	7,426	9,259	2,121	8,242	7,536
Original	Logistic Regression	7,193	7,475	9,136	2,424	8,222	8,036
SMOTENC	Random Forest	7,281	7,841	8,519	4,242	8,166	7,727
Original	KNN	7,105	7,553	8,765	3,030	8,114	7,024
SMOTENC	SVM	7,456	9,063	7,160	8,182	8,000	8,275
SMOTENC	Logistic Regression	7,456	9,194	7,037	8,485	7,972	8,257

The best-performing model based on testing F1-score was Gradient Boosting with SMOTENC, achieving an accuracy of 0.7632, precision of 0.7935, recall of 0.9012, specificity of 0.4242, F1-score of 0.8439, and ROC-AUC of 0.7759. This result indicates that the model was effective in identifying liver disease cases, as reflected by its high recall value.

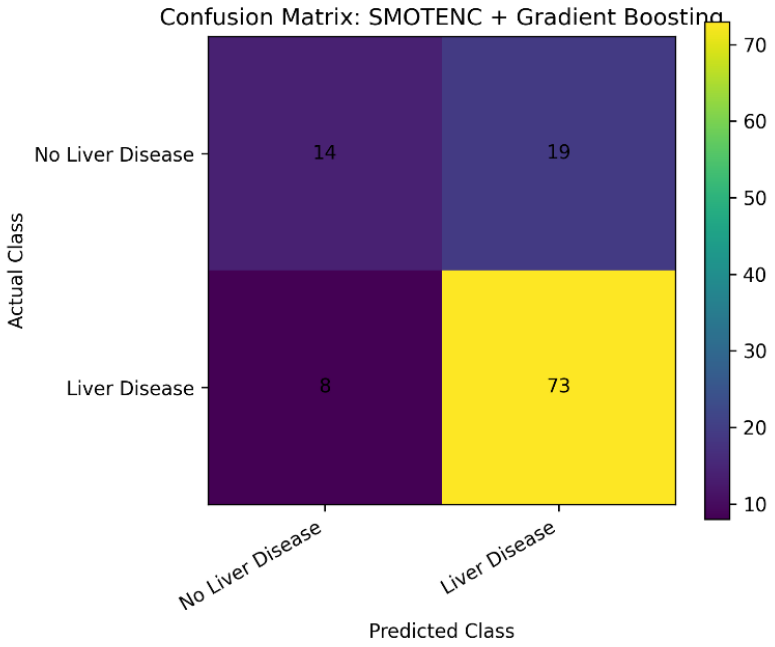

Figure 5. Best Model

The confusion matrix of the best model is shown in **Figure 5**. The model correctly classified 73 liver disease patients and 14 non-liver disease patients. However, it misclassified 19 non-liver disease patients as liver disease and 8 liver disease patients as no liver disease. These results show that the model had a strong ability to detect liver disease cases, but its ability to correctly recognize non-liver disease cases was more limited.

Table 6. Confusion Matrix of the Best Model

Actual Class	**Predicted No Liver Disease**	**Predicted Liver Disease**
No Liver Disease	14	19
Liver Disease	8	73

From a clinical screening perspective, the high recall value is important because it means the model can identify most patients with liver disease. In this study, only 8 out of 81 liver disease cases in the testing set were missed by the best model. However, the relatively low specificity indicates that the model still produced false positive predictions among patients without liver disease. This means that the model is more suitable as an early screening support tool rather than as a standalone diagnostic system [24].

Effect of Class Imbalance Handling:

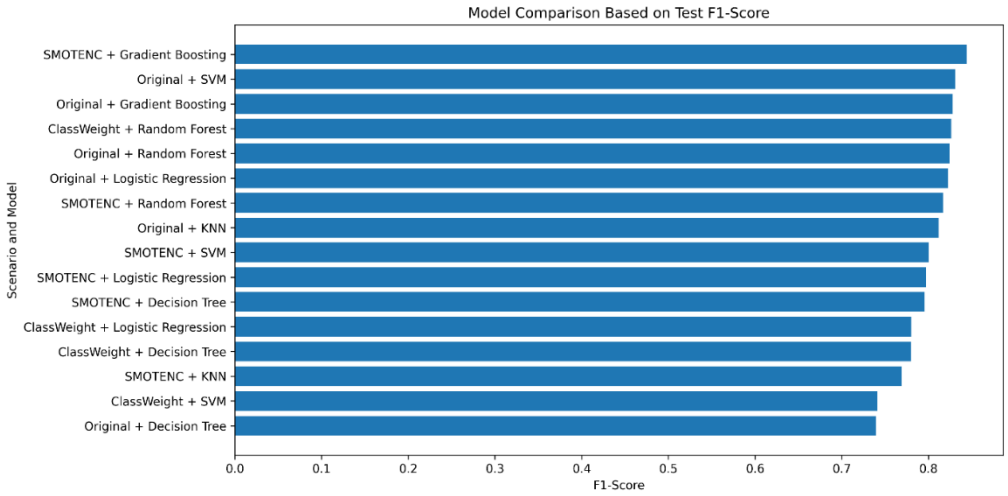

Figure 6. Model Comparison Based on Test F1-Score

The comparison of F1-score across models is presented in **Figure 6**. The results show that SMOTENC combined with Gradient Boosting produced the highest F1-score. This suggests that oversampling the minority class helped improve the balance between precision and recall for the selected model.

However, the effect of imbalance handling was not uniform across all algorithms. Some models trained on the original dataset, such as SVM and Gradient Boosting, also achieved high F1-scores [25]. Nevertheless, the original SVM model achieved a recall of 1.0000 but a specificity of 0.0000, meaning that it classified all testing samples as liver disease. Although this produced a high F1-score due to the dominance of liver disease cases in the dataset, it failed to identify any patient without liver disease. This finding demonstrates why accuracy and F1-score should not be interpreted alone in imbalanced medical datasets

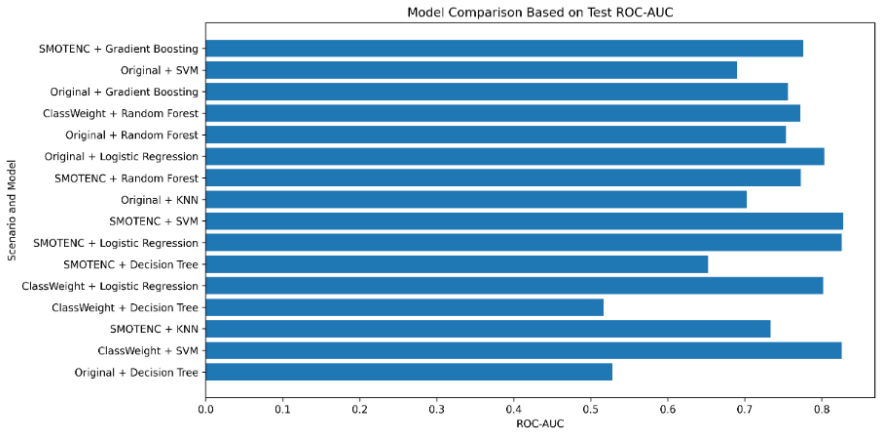

Figure 7. Model Comarison Based on ROC AUC.

The ROC-AUC comparison is shown in Figure 7. The highest ROC-AUC was obtained by SMOTENC with SVM, with a value of 0.8275, followed closely by SMOTENC with Logistic Regression and ClassWeight with SVM, both achieving ROC-AUC values above 0.82. These models also produced higher specificity than the best F1-score model. For example, SMOTENC with Logistic Regression achieved a specificity of 0.8485, while SMOTENC with SVM achieved a specificity of 0.8182. However, their recall values were lower than Gradient Boosting with SMOTENC.

This result indicates a trade-off between sensitivity and specificity. Gradient Boosting with SMOTENC was more effective for detecting liver disease cases, while SMOTENC-based Logistic Regression and SVM were better at reducing false positives among non-liver disease cases. In the context of early liver disease screening, higher sensitivity is often preferable because missing a potential liver disease case may delay further medical evaluation. Therefore, Gradient Boosting with SMOTENC was selected as the best model in this study based on its overall F1-score and strong sensitivity.

Gender-Based Performance Analysis:

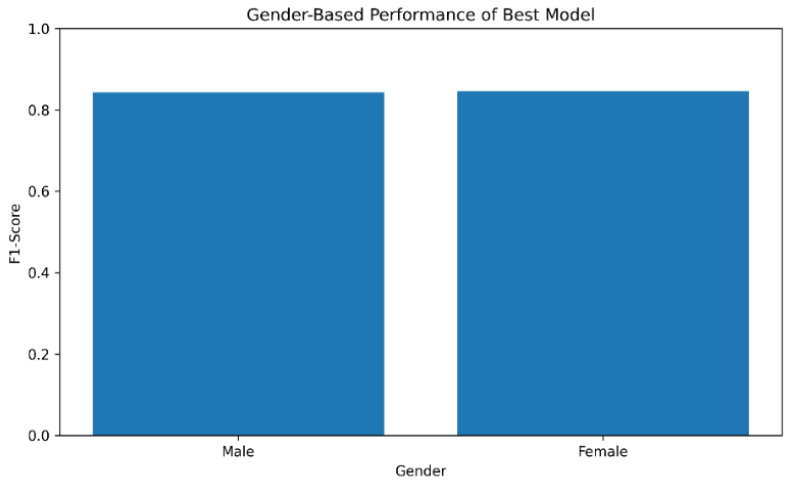

Figure 8. Gender-Based Performance of Best Model

The gender-based performance of the best model is shown in Figure 8 and summarized in Table 7. The testing set contained 81 male patients and 33 female patients. The model achieved similar F1-scores for male and female patients, with an F1-score of 0.8430 for male patients and 0.8462 for female patients.

Table 7. Gender-Based Performance of the Best Model

Actual Class	Predicted No Liver Disease	Predicted Liver Disease
No Liver Disease	14	19
Liver Disease	8	73

The results show that the model performed consistently across gender groups in terms of F1-score. The difference between male and female F1-scores was very small, suggesting that the model did not show a major performance gap between the two groups. The recall value was slightly higher for male patients, while precision was slightly higher for female patients.

However, specificity remained low in both gender groups, especially among female patients. The model achieved a specificity of 0.4400 for male patients and 0.3750 for female patients. This indicates that the model was less effective in correctly identifying patients without liver disease in both groups. The lower specificity in female patients should be interpreted carefully because the number of female samples in the testing set was relatively small.

Overall, the gender-based evaluation suggests that the best model produced similar F1-score performance for male and female patients. Nevertheless, the unequal gender distribution in the dataset and the relatively small number of female testing samples limit the strength of fairness-related conclusions. Future studies should use a larger and more balanced dataset to validate whether the model remains stable across gender groups.

Feature Importance Analysis:

Feature importance analysis was conducted using permutation importance to identify which variables contributed most strongly to the best model. The results are shown in **Figure 9** and summarized in **Table 8**.

Table 8. Feature Importance of the Best Model

Gender	N	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC
Male	81	7,654	7,846	9,107	4,400	8,430	7,907
Female	33	7,576	8,148	8,800	3,750	8,462	7,450

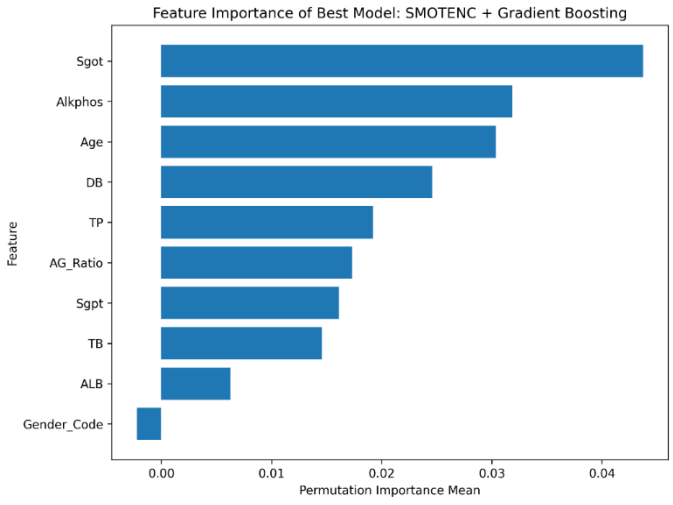

Figure 9. Feature Importance of the Best Model

The most influential feature was SGOT, followed by alkaline phosphatase, age, and direct bilirubin. These results are clinically meaningful because SGOT, alkaline phosphatase, and bilirubin-related markers are commonly associated with liver function and liver injury. The importance of age also suggests that demographic characteristics contribute to the risk pattern captured by the model [26].

Interestingly, gender showed a very small negative permutation importance value. This indicates that gender did not contribute meaningfully to improving the predictive performance of the best model. This finding is consistent with the gender-based performance analysis, where the model showed relatively similar F1-scores for male and female patients. However, this does not mean that gender is clinically irrelevant. Rather, in this specific dataset and modeling configuration, laboratory biomarkers were more informative than gender for predicting liver disease status.

The feature importance results improve the interpretability of the model. Instead of functioning as a black-box classifier, the model provides insight into which clinical variables were most influential in the prediction process. This is important for medical artificial intelligence because interpretability can help clinicians and researchers understand whether the model relies on clinically reasonable variables.

Discussion:

The findings of this study demonstrate that machine learning can be used to predict liver disease based on demographic and biochemical clinical features [27]. Among the evaluated models, Gradient Boosting combined with SMOTENC achieved the best F1-score and showed strong sensitivity in detecting liver disease cases. This suggests that ensemble-based learning combined with imbalance handling can improve predictive performance in liver disease classification.

The high recall of the best model is particularly relevant for early screening. In healthcare applications, especially for disease risk identification, failing to detect a patient with a potential disease may have serious consequences. Therefore, a model with high sensitivity can be valuable as an initial decision-support tool. The model can help identify patients who may require further medical examination, laboratory confirmation, or specialist consultation.

However, the model's relatively low specificity indicates that it still incorrectly classified several non-liver disease patients as liver disease. In clinical practice, this may lead to unnecessary follow-up examinations. While false positives are generally less harmful than false negatives in early screening contexts, they still need to be minimized to improve clinical efficiency. Therefore, the selected model should not be interpreted as a definitive diagnostic tool, but rather as a screening-support model.

The results also highlight the importance of using multiple evaluation metrics in imbalanced medical datasets. For example, the original SVM model achieved a high F1-score and perfect recall, but its specificity was zero. This means the model failed completely in identifying non-liver disease cases. If only F1-score or recall were considered, the model might appear strong. However, the specificity result shows that it was not clinically reliable. This confirms that medical AI models should be evaluated using a broader set of metrics, including sensitivity, specificity, ROC-AUC, and confusion matrix analysis.

The gender-based evaluation showed that the best model achieved similar F1-scores for male and female patients. This suggests that the model's overall classification performance was relatively stable across gender groups. Nevertheless, the dataset contained far more male than female patients, and the female testing subset was small. Therefore, the gender-aware analysis should be interpreted as an exploratory subgroup evaluation rather than a definitive fairness assessment [28].

The feature importance analysis showed that SGOT, alkaline phosphatase, age, and direct bilirubin were the most influential variables in the prediction process. These findings are consistent with the clinical relevance of liver enzyme and bilirubin markers. The relatively low importance of gender suggests that biochemical markers contributed more strongly to the model's predictions than demographic gender information.

Overall, this study contributes to AI-based liver disease prediction by combining model comparison, class imbalance handling, subgroup-level evaluation, and explainability analysis. The proposed approach provides a more comprehensive evaluation framework than a simple accuracy-based comparison. The results indicate that an explainable and gender-aware machine learning pipeline can support liver disease screening using routinely available clinical biomarkers [29].

Limitations:

This study has several limitations. First, the dataset size was relatively small, with only 570 records after duplicate removal. Second, the dataset was imbalanced both in terms of target class and gender distribution. Third, the data came from a specific regional population, which may limit generalization to other populations or healthcare settings. Fourth, the model was developed using structured laboratory data only and did not include additional clinical information such as symptoms, medical history, hepatitis status, alcohol consumption, obesity, imaging results, or physician diagnosis notes.

Therefore, future studies should validate the proposed approach using larger, more diverse, and multi-center datasets. Additional clinical variables may also be included to improve prediction performance and clinical relevance. Furthermore, advanced explainability methods such as SHAP may be applied to provide more detailed individual-level interpretation of model predictions [30].

4. Conclusion

This study developed and evaluated an explainable gender-aware machine learning framework for liver disease prediction using demographic and clinical biomarker data. The dataset consisted of 570 patient records after duplicate removal, with liver disease cases representing the majority class. The analysis showed that the dataset was imbalanced not only in terms of target class distribution, but also in gender distribution, where male patients were more dominant than female patients. These characteristics justified the use of class imbalance handling and subgroup-based performance evaluation.

Several machine learning algorithms were compared under three experimental scenarios, namely original data, class-weighted learning, and SMOTENC-based oversampling. The experimental results showed that Gradient

Boosting combined with SMOTENC produced the best performance based on the testing F1-score. This model achieved an accuracy of 0.7632, precision of 0.7935, recall or sensitivity of 0.9012, specificity of 0.4242, F1-score of 0.8439, and ROC-AUC of 0.7759. The high sensitivity indicates that the model was effective in identifying patients with liver disease, making it potentially useful as an early screening support tool.

The confusion matrix analysis showed that the best model correctly identified 73 out of 81 liver disease patients in the testing set. However, the model also misclassified 19 non-liver disease patients as liver disease. This finding indicates that although the model had strong sensitivity, its specificity remained limited. Therefore, the proposed model should not be used as a standalone diagnostic system, but rather as a decision-support tool to assist early risk identification and recommend further clinical examination.

The gender-based evaluation showed that the best model achieved similar F1-score performance for male and female patients, with an F1-score of 0.8430 for male patients and 0.8462 for female patients. This suggests that the model demonstrated relatively stable predictive performance across gender groups. However, because the dataset contained a smaller number of female patients, this result should be interpreted as an exploratory subgroup analysis rather than a definitive fairness assessment.

Feature importance analysis revealed that SGOT, alkaline phosphatase, age, and direct bilirubin were the most influential variables in predicting liver disease. These findings are clinically reasonable because liver enzymes and bilirubin-related markers are commonly associated with liver function abnormalities. The relatively low importance of gender suggests that biochemical markers contributed more strongly to model prediction than demographic gender information in this dataset.

Overall, this study demonstrates that machine learning can support liver disease prediction using routinely available clinical biomarkers. The integration of class imbalance handling, gender-based evaluation, and feature importance analysis provides a more comprehensive framework for developing clinically relevant AI-based screening models. Future research should validate the proposed approach using larger, more balanced, and multi-center datasets. Additional clinical variables, such as symptoms, medical history, hepatitis status, alcohol consumption, obesity indicators, and imaging findings, should also be considered to improve predictive performance and clinical applicability.

References:

[1] H. Devarbhavi, S. K. Asrani, J. P. Arab, Y. A. Nartey, E. Pose, and P. S. Kamath, “Global burden of liver disease: 2023 update,” *J. Hepatol.*, vol. 79, no. 2, pp. 516–537, Aug. 2023, doi: 10.1016/j.jhep.2023.03.017.

[2] X.-N. Wu *et al.*, “Global burden of liver cirrhosis and other chronic liver diseases caused by specific etiologies from 1990 to 2019,” *BMC Public Health*, vol. 24, no. 1, p. 363, Feb. 2024, doi: 10.1186/s12889-024-17948-6.

[3] S. Xiao, W. Xie, Y. Zhang, L. Lei, and Y. Pan, “Changing epidemiology of cirrhosis from 2010 to 2019: results from the Global Burden Disease study 2019,” *Ann. Med.*, vol. 55, no. 2, p. 2252326, Dec. 2023, doi: 10.1080/07853890.2023.2252326.

[4] S. Thakur, V. Kumar, R. Das, V. Sharma, and D. K. Mehta, "Biomarkers of Hepatic Toxicity: An Overview," *Curr. Ther. Res.*, vol. 100, p. 100737, 2024, doi: 10.1016/j.curtheres.2024.100737.

[5] R. A. Khan, Y. Luo, and F.-X. Wu, "Machine learning based liver disease diagnosis: A systematic review," *Neurocomputing*, vol. 468, pp. 492–509, Jan. 2022, doi: 10.1016/j.neucom.2021.08.138.

[6] S. M. Ganie, P. K. Dutta Pramanik, and Z. Zhao, "Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 160, Jun. 2024, doi: 10.1186/s12911-024-02550-y.

[7] A. U. Rehman *et al.*, "A Machine Learning-Based Framework for Accurate and Early Diagnosis of Liver Diseases: A Comprehensive Study on Feature Selection, Data Imbalance, and Algorithmic Performance," *Int. J. Intell. Syst.*, vol. 2024, no. 1, p. 6111312, Jan. 2024, doi: 10.1155/2024/6111312.

[8] W. El Atifi, O. El Rhazouani, F. M. Khan, and H. Sekkat, "Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare," *PLOS One*, vol. 20, no. 8, p. e0330899, Aug. 2025, doi: 10.1371/journal.pone.0330899.

[9] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, p. 581, Feb. 2023, doi: 10.3390/biomedicines11020581.

[10] S. Dalal, E. M. Onyema, and A. Malik, "Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy," *World J. Gastroenterol.*, vol. 28, no. 46, pp. 6551–6563, Dec. 2022, doi: 10.3748/wjg.v28.i46.6551.

[11] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Health*, vol. 6, p. 1430245, Jul. 2024, doi: 10.3389/fdgth.2024.1430245.

[12] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," *Inform. Med. Unlocked*, vol. 36, p. 101155, 2023, doi: 10.1016/j.imu.2022.101155.

[13] S. K. Joo and W. Kim, "Sex differences in metabolic dysfunction-associated steatotic liver disease: a narrative review," *Ewha Med. J.*, vol. 47, no. 2, p. e17, Apr. 2024, doi: 10.12771/emj.2024.e17.

[14] I. Straw and H. Wu, "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction," *BMJ Health Care Inform.*, vol. 29, no. 1, p. e100457, Apr. 2022, doi: 10.1136/bmjhci-2021-100457.

[15] Q. Abbas, W. Jeong, and S. W. Lee, "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges," *Healthcare*, vol. 13, no. 17, p. 2154, Aug. 2025, doi: 10.3390/healthcare13172154.

[16] R. Rani *et al.*, "Enhancing liver disease diagnosis with hybrid SMOTE-ENN balanced machine learning models—an empirical analysis of Indian patient liver disease datasets," *Front. Med.*, vol. 12, p. 1502749, May 2025, doi: 10.3389/fmed.2025.1502749.

[17] B. Njei, E. Osta, N. Njei, Y. A. Al-Ajlouni, and J. K. Lim, "An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis," *Sci. Rep.*, vol. 14, no. 1, p. 8589, Apr. 2024, doi: 10.1038/s41598-024-59183-4.

[18] J. Deng *et al.*, "Development and validation of a machine learning-based framework for assessing metabolic-associated fatty liver disease risk," *BMC Public Health*, vol. 24, no. 1, p. 2545, Sep. 2024, doi: 10.1186/s12889-024-19882-z.

[19] F. Masaebi *et al.*, "Machine-Learning Application for Predicting Metabolic Dysfunction-Associated Steatotic Liver Disease Using Laboratory and Body Composition Indicators," *Arch. Iran. Med.*, vol. 27, no. 10, pp. 551–562, Oct. 2024, doi: 10.34172/aim.31269.

[20] L. Zhang, Y. Huang, M. Huang, C.-H. Zhao, Y.-J. Zhang, and Y. Wang, "Development of Cost-Effective Fatty Liver Disease Prediction Models in a Chinese Population: Statistical and Machine Learning Approaches," *JMIR Form. Res.*, vol. 8, p. e53654, Feb. 2024, doi: 10.2196/53654.

[21] B. Yang, H. Lu, and Y. Ran, "Advancing non-alcoholic fatty liver disease prediction: a comprehensive machine learning approach integrating SHAP interpretability and multi-cohort validation," *Front. Endocrinol.*, vol. 15, p. 1450317, Oct. 2024, doi: 10.3389/fendo.2024.1450317.

[22] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, Jun. 2024, doi: 10.1016/j.patter.2024.100994.

[23] M. Salimparsa, K. Sedig, D. J. Lizotte, S. S. Abdullah, N. Chalabianloo, and F. T. Muanda, "Explainable AI for Clinical Decision Support Systems: Literature Review, Key Gaps, and Research Synthesis," *Informatics*, vol. 12, no. 4, p. 119, Oct. 2025, doi: 10.3390/informatics12040119.

[24] M. Pons *et al.*, "Point-of-Care Noninvasive Prediction of Liver-Related Events in Patients With Nonalcoholic Fatty Liver Disease," *Clin. Gastroenterol. Hepatol.*, vol. 22, no. 8, pp. 1637-1645.e9, Aug. 2024, doi: 10.1016/j.cgh.2023.08.004.

[25] V. Charu, J. W. Liang, A. Mannalithara, A. Kwong, L. Tian, and W. R. Kim, "Benchmarking clinical risk prediction algorithms with ensemble machine learning for the noninvasive diagnosis of liver fibrosis in NAFLD," *Hepatology*, vol. 80, no. 5, pp. 1184–1195, Nov. 2024, doi: 10.1097/HEP.0000000000000908.

[26] Y. Yu, Y. Yang, Q. Li, J. Yuan, and Y. Zha, "Predicting metabolic dysfunction associated steatotic liver disease using explainable machine learning methods," *Sci. Rep.*, vol. 15, no. 1, p. 12382, Apr. 2025, doi: 10.1038/s41598-025-96478-6.

[27] C.-H. Lu *et al.*, “Machine Learning Models for Predicting Significant Liver Fibrosis in Patients with Severe Obesity and Nonalcoholic Fatty Liver Disease,” *Obes. Surg.*, vol. 34, no. 12, pp. 4393–4404, Dec. 2024, doi: 10.1007/s11695-024-07548-z.

[28] A. Talwar *et al.*, “Sex bias consideration in healthcare machine-learning research: a systematic review in rheumatoid arthritis,” *BMJ Open*, vol. 15, no. 3, p. e086117, Mar. 2025, doi: 10.1136/bmjopen-2024-086117.

[29] S. Weng, D. Hu, J. Chen, Y. Yang, and D. Peng, “Prediction of Fatty Liver Disease in a Chinese Population Using Machine-Learning Algorithms,” *Diagnostics*, vol. 13, no. 6, p. 1168, Mar. 2023, doi: 10.3390/diagnostics13061168.

[30] N. Almusallam and S. Khan, “Chronic liver disease classification using deep learning with SHAP-optimized hybrid features,” *iScience*, vol. 28, no. 12, p. 113972, Dec. 2025, doi: 10.1016/j.isci.2025.113972.