Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 4 (2026) 1025 - 1033 e-ISSN: 2580-0760

LightGBM for Liver Disease Detection with Hybrid Hyperparameter Optimization

Fajar Ratnawati[1], Agus Tedyyana[2], Johny Custer[3]
[1,2]Department of Informatic Engineering, Politeknik Negeri Bengkalis, Indonesia
[3]Department of Electrical Engineering, Politeknik Negeri Bengkalis, Indonesia

[*]*fajar@polbeng.ac.id*

Abstract

In response to the growing burden of liver-related disorders, this research develops a supervised learning approach using the Light Gradient Boosting Machine (LightGBM) algorithm to support the early identification of Non-Alcoholic Fatty Liver Disease (NAFLD). The study focuses on constructing and assessing a robust classification model that differentiates individuals with NAFLD from those without the condition based on routinely collected clinical indicators and lifestyle-related characteristics. The dataset, obtained from an open-access NAFLD repository, consists of 1,700 patient records with 10 predictor variables and one binary diagnosis label. The proposed framework employs a stratified shuffle split evaluation scheme with 5-fold and 10-fold cross-validation, using out-of-fold (OOF) probabilities to compute overall performance metrics. The baseline LightGBM model already demonstrated strong performance, achieving 88.94% accuracy, 90.74% precision, 88.99% recall, 89.85% F1-score, and 92.52% AUC under 10-fold cross-validation. To further improve predictive performance, hyperparameter tuning was performed using Optuna and Bayesian Optimization. Among the evaluated approaches, Bayesian-optimized LightGBM achieved the best results, with 93.17% accuracy, 94.49% precision, 92.52% recall, 93.72% F1-score, and 93.28% AUC under 10-fold cross-validation. These findings indicate that systematic hyperparameter optimization can improve the discriminative capability of LightGBM for NAFLD detection and support its potential as a reliable decision-support tool in clinical settings.

Keywords: bayesian optimization; hyperparameter optimization; LightGBM; liver disease; NAFLD; Optuna

How to Cite: F. Ratnawati, A. Tedyyana, and J. Custer, "LightGBM for Liver Disease Detection with Hybrid Hyperparameter Optimization", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 4, pp. 1025 - 1033, Aug. 2026.
*Permalink/*DOI: https://doi.org/10.29207/resti.v10i4.7357

Received: November 24, 2025
Accepted: June 21, 2026
Available Online: August 10, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The rapid advancement of information technology and the increasing availability of large-scale health data have driven the development of various intelligent computational approaches to support clinical decision-making [1]. With non-communicable diseases on the rise around the world, machine learning has become a promising way to find them early, figure out who is at risk, and predict how they will progress. Liver disease comprises a broad spectrum of disorders that have received considerable medical and research attention. These conditions range from metabolic dysfunction–associated steatotic liver disease (MASLD) and viral hepatitis to more advanced stages, including cirrhosis and hepatocellular carcinoma. MASLD is now known to be the most common liver disease in the world, affecting about 25–38% of the global population. The number of people with it is rising along with the number of people who are obese and have type 2 diabetes. This condition implies that the at-risk population for liver disease is actually enormous, yet often remains undiagnosed because the disease is frequently asymptomatic in its early stages [2].

The burden of liver disease becomes even more concerning when viewed from the perspective of global morbidity and mortality. According to GLOBOCAN 2020, liver cancer was one of the major contributors to global cancer mortality, accounting for an estimated 905,677 newly diagnosed cases and 830,180 deaths during that year. The high mortality-to-incidence ratio indicates that many cases are diagnosed at advanced stages, when treatment options are limited, and the likelihood of cure decreases sharply. A similar pattern is observed in Indonesia, where GLOBOCAN 2020

shows that liver cancer belongs to the group of cancers with high incidence and mortality, with tens of thousands of new cases and deaths annually, positioning it as a serious challenge for the national health system [3].

In parallel, several studies in the health domain have demonstrated that gradient boosting family algorithms, particularly Light Gradient Boosting Machine (LightGBM), can deliver highly competitive predictive performance. For example, one study employed LightGBM to predict insulin resistance based on a combination of clinical and metabolomic features and found that LightGBM outperformed other models in terms of accuracy and its ability to capture non-linear relationships among metabolic risk factors [4].

These findings suggest that LightGBM is highly promising for application to liver disease prediction, given the similar data characteristics, namely multiple interacting clinical and biochemical variables. However, the use of LightGBM specifically for liver disease detection, especially within rigorously designed, early-detection-oriented experimental settings, remains relatively limited in the literature, particularly in the context of populations and datasets relevant to the Indonesian setting. Hyperparameter optimization strongly influences the quality of predictive models beyond algorithm selection.

Many previous studies still rely on simple approaches such as grid search or manual trial-and-error, which are time-consuming, prone to bias, and often explore only a narrow region of the hyperparameter space [5]. Recently, more advanced and efficient hyperparameter optimization frameworks have emerged, one of which is Optuna, an open-source library that implements the Tree-structured Parzen Estimator (TPE) approach, automatic pruning, and adaptive search, enabling the discovery of good hyperparameter configurations with fewer trials compared to grid or random search. Optuna has been widely applied to tune state-of-the-art machine learning models, including gradient boosting methods [6].

In parallel, Bayesian optimization has long been recognized as a powerful and sample-efficient approach for black-box optimization, including the tuning of machine learning hyperparameters. Several studies in the health domain have applied Bayesian optimization to tune hyperparameters of disease classification models, for example, in predicting HIV infection risk and in modeling survival and progression of other chronic diseases, and have reported consistent performance gains over conventional tuning strategies [7].

Bayesian optimization works by probabilistically modeling the relationship between hyperparameter configurations and model performance (e.g., AUC) using surrogate models such as Gaussian processes or tree-based models, and then iteratively selecting new candidate points that are expected to be most informative [8]. Nevertheless, most of these studies focus on a single optimization framework (e.g., only Bayesian Optimization or only TPE/Optuna), without exploring the potential of combining multiple optimization strategies in a hybrid manner to maximize model performance.

From this review, several important research gaps can be identified. First, although there are existing studies that apply machine learning algorithms for liver disease detection, most still rely on classical methods (Naive Bayes, K-Nearest Neighbors, Decision Tree, and Random Forest) whose performance is not yet optimal and which do not deeply explore next-generation gradient boosting algorithms such as LightGBM that are theoretically and empirically superior for handling high-dimensional and non-linear data[9].

Second, although Optuna and Bayesian optimization have each been shown to be effective for hyperparameter optimization in various contexts, their combined use within a hybrid optimization scheme to refine LightGBM models for liver disease detection has not been widely investigated. Most prior studies simply compare a single optimization technique with a basic baseline and therefore do not systematically address whether a hybrid approach can yield significant and stable performance improvements.

Motivated by these gaps, this study proposes a LightGBM-based model for liver disease detection and evaluates two hyperparameter optimization approaches, namely Optuna and Bayesian Optimization, to obtain more optimal model configurations than those produced by a single tuning strategy alone. Empirically, the study compares the predictive performance of the baseline LightGBM model, the Optuna-optimized LightGBM model, and the Bayesian-optimized LightGBM model using comprehensive evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. To ensure a more reliable assessment and minimize data-splitting bias, the evaluation was conducted using a Stratified Shuffle Split scheme.

2. Research Methods

Our study used a supervised machine learning approach to develop a model for assessing the risk of liver disease based on a cross-sectional dataset. The dataset, sourced from the public repository Kaggle (https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset), consists of records from 1,700 patients undergoing liver disease screening. The data collection protocol encompassed demographic data, behavioral risk factors, medical history, and diagnostic test outcomes, thereby constructing a thorough representation of each patient's risk profile across various domains of potential carcinogenic impact. The ethical framework governing this study prioritized patient privacy; all records were de-identified before analysis using standard medical research protocols. Methodologically, as illustrated in Figure 1, this study proposes a framework for

developing a LightGBM-based model that is combined with two hyperparameter optimization approaches Optuna (TPE-based) and Bayesian Optimization, each of which has distinct characteristics in exploring the hyperparameter space.

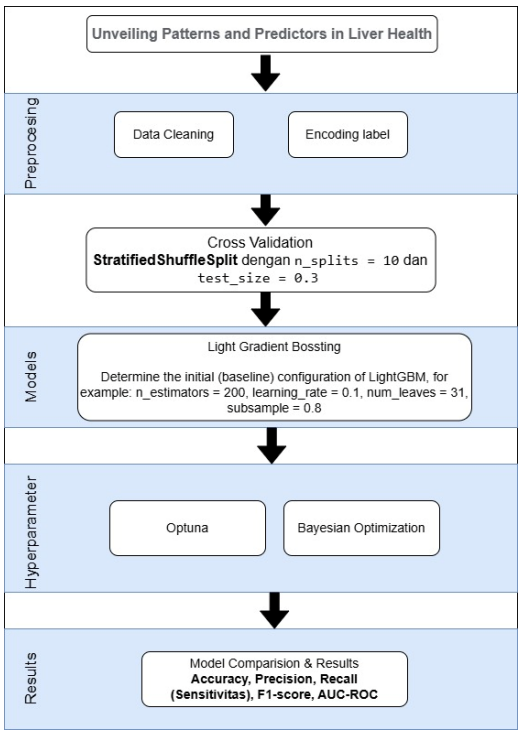


Figure 1. Research Flow

LightGBM is a tool created by Microsoft Research Asia that uses the Gradient Boosting Decision Tree (GBDT) method to make it faster and easier to solve prediction problems with large amounts of data. Relative to alternative GBDT frameworks, LightGBM provides several computational and predictive benefits. These include shorter training times, efficient resource utilization, reduced memory requirements, competitive predictive performance, scalability to large datasets, and compatibility with both parallel processing and GPU-accelerated training [10]. LightGBM is a quick and powerful tool that uses decision tree algorithms for gradient boosting, and it can be used for tasks like ranking, classification, regression, and other machine learning applications [11]. The core concept can be represented by the following Equation 1.

$$f(x) = \sum_{q=1}^{Q} \alpha_q \mathrm{T}(\mathrm{x}, \theta_\mathrm{q}) \quad (1)$$

In Equation (1), f(x) denotes the final prediction produced by the LightGBM ensemble for the input feature vector x. Q is the total number of decision-tree learners, and q is the index of an individual learner. The coefficient αq represents the contribution (or weight) of the q-th learner, θq denotes the parameters of that learner, and T(x; θq) is the output generated by the q-th decision tree for input x. Thus, the final prediction is obtained as the weighted sum of the outputs of all Q base learners. Once the training data and loss function have been specified, the boosting training procedure can be formulated as an optimization task that minimizes the empirical loss between the observed target values and the predictions generated by the ensemble model, as expressed in Equation 2 [12].

$$\sum_{i=1}^{N} (y_i, \hat{y}_i) \quad (2)$$

(N) denotes the total number of training samples, (i) refers to the index of an individual sample, (y_i) represents the observed target value of the (i)-th sample, and (ˆ{y}_i) denotes the value predicted by the LightGBM ensemble model. The function $\big(l(y_i, \hat{y}_i)\big)$ represents the loss incurred for the $(i) - th$ observation. Therefore, the objective function minimizes the aggregate prediction error across all training samples.

Y_h represents the observed value of the $h - th$ sample, whereas $f(X_h)$ denotes the value predicted for that sample. The term $L\big(Y_h, f(X_h)\big)$ expresses the loss associated with the $h - th$ observation. In Gradient Boosting Machine (GBM), successive decision-tree models are added iteratively using a gradient-descent procedure. At each iteration, the newly introduced sub-model is trained to reduce the residual error by moving in the direction of the negative gradient of the loss function, thereby progressively minimizing the overall prediction error [13], as shown in Equation 3.

$$\arg min \sum_{h=1}^{H} L\big(Y_h, f(X_h)\big) \quad (3)$$

$L(F_j(x), (Y)$ and $L\big(F_j - 1(x), Y\big)$ represent the loss values obtained at the $j - th$ and $(j - 1)$th iterations, respectively. Likewise, $F_j(x)$ and $F_j - 1(x)$ denote the model predictions generated at those iterations Y represents the actual target value of the sample as shown in Equation 4 [14].

$$L(F_j(x), (Y) < L(F_j - 1(x), Y) \quad (4)$$

A distinctive feature of LightGBM compared with other gradient-boosted decision-tree algorithms is its leaf-wise tree construction mechanism. Rather than expanding all nodes at an identical depth simultaneously, as performed in the conventional level-wise approach, LightGBM selects and splits the leaf that provides the greatest decrease in the loss function. This strategy can accelerate loss minimization and improve predictive performance, although the tree depth must be properly controlled to prevent overfitting [15] A comparison between the leaf-wise and level-wise tree growth mechanisms is presented in Figure 2.

Optuna is utilized to optimize model hyperparameters through a Bayesian search mechanism implemented using the Tree-structured Parzen Estimator (TPE). During the optimization process, information obtained from earlier trials is used to estimate which areas of the search space are more likely to produce better results. This adaptive procedure allows subsequent evaluations to focus on promising hyperparameter combinations rather than examining all possible configurations uniformly. Consequently, Optuna can identify suitable

parameter settings with fewer trials and lower computational cost than exhaustive methods such as grid search [16].

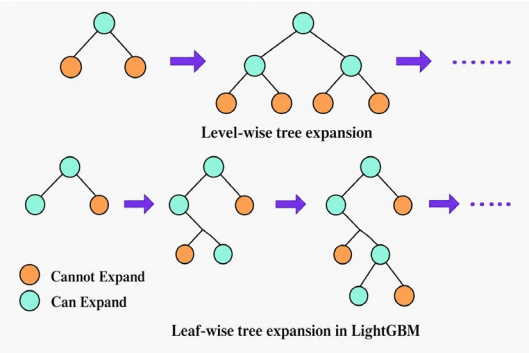


Figure 2. LightGBM

Bayesian optimization is a probabilistic optimization method used to search for the best combination of parameter values for a model with as few iterations as possible. In this study, Bayesian optimization is employed to tune the hyperparameters of the LightGBM model. The optimization process proceeds iteratively through two main components, namely a surrogate model and an acquisition function, which together guide the selection of the most promising hyperparameter configurations at each step [17].

Stratified Shuffle Split is a data splitting technique that partitions a dataset into training and testing sets in a random manner while preserving the original class distribution in each subset. In other words, the proportion of each label (e.g., liver disease vs. non–liver disease) is kept approximately the same in both the training and testing data. This procedure can be repeated multiple times to generate several randomized yet class-balanced splits. By maintaining a proportional representation of all classes, Stratified Shuffle Split helps ensure that the model is trained and evaluated fairly across all categories, reducing the risk of bias toward majority classes and improving the reliability of performance estimates[18].

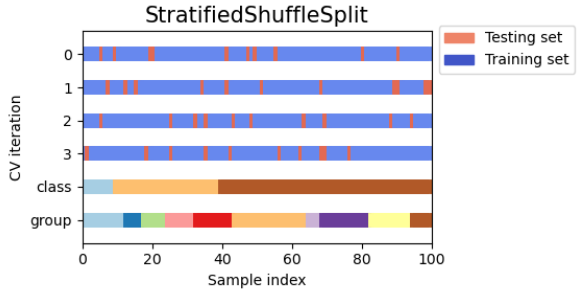


Figure 3. Stratified Shuffle Split

3. Results and Discussions

3.1 Results

The dataset used in this study comprises 1,700 patient records with 10 predictor variables, including age, gender, body mass index (BMI), alcohol consumption, smoking status, genetic risk, physical activity, diabetes, hypertension, and liver function test results. The diagnosis column is treated as the target variable, representing whether a patient is classified as having liver disease (class 1) or not (class 0). These variables were selected because they reflect both lifestyle-related factors and clinical indicators commonly associated with liver disease risk.

In terms of class composition, the dataset is fairly balanced, containing 764 samples in class 0 and 936 samples in class 1. This means that approximately 55% of the patients are labeled as having liver disease. Although class 1 is slightly more frequent, the imbalance is mild, which helps reduce the likelihood that the model will overly favor one class during training and evaluation. This distribution also supports the use of standard classification metrics without requiring aggressive rebalancing strategies.

To ensure a more reliable and less split-dependent evaluation, the study adopts an Out-of-Fold (OOF) evaluation scheme using Stratified Shuffle Split, which preserves class proportions in each split. Experiments are conducted under 5-fold and 10-fold settings with a 30% test proportion per split. In each fold, the model is trained on the stratified training subset and then produces predicted probabilities for the corresponding test subset; these outputs are accumulated across folds to form complete OOF predictions, ensuring each record is evaluated by a model that did not see it during training. Finally, performance is reported using accuracy, precision, recall, F1-score, and AUC-ROC, providing a comprehensive view of both classification correctness and discriminative ability.

Table 1. Attributes Description

Attributes	Description
Age	Patient age, ranging from 20 to 80 years.
Gender	Patient gender: male (0) or female (1).
BMI	Body mass index, ranging from 15 to 40.
AlcoholConsumption	Weekly alcohol intake, ranging from 0 to 20 units.
Smoking	Smoking status: no (0) or yes (1).
GeneticRisk	Genetic predisposition to liver disease: low (0), medium (1), or high (2).
PhysicalActivity	Duration of physical activity per week, ranging from 0 to 10 hours.
Diabetes	Diabetes status: no (0) or yes (1).
Hypertension	Hypertension status: no (0) or yes (1).
LiverFunctionTest	Liver function test value, ranging from 20 to 100.
Diagnosis	Target variable indicating the absence (0) or presence (1) of liver disease.

Table 1 summarizes the dataset structure, which comprises ten independent variables and one dependent variable. The input features cover four main categories: demographic information, lifestyle-related factors, prior health conditions, and clinical indicators. Continuous or numerical attributes include age, body mass index, alcohol intake, weekly physical activity, and liver function test values. In contrast, gender, smoking behavior, diabetes status, and hypertension status are represented using binary encoding. Genetic

risk is treated as an ordinal variable with three levels, namely low, moderate, and high. The diagnosis variable is defined as the classification target, with a value of 0 representing individuals without liver disease and a value of 1 identifying those diagnosed with the condition.

In this study, the dataset testing used the LightGBM model had parameters in Table 2, with cross validation of 5 and 10, so that the accuracy results were obtained as in Table 3.

Table 2. LightGBM Parameters (Base)

No	Parameter	Value
1	boosting_type	"gbdt"
2	objective	"binary"
3	metric	"binary_logloss"
4	n_estimators	200
5	learning_rate	0.1
6	num_leaves	31
7	max_depth	-1
8	subsample	0.8
9	colsample_bytree	0.8
10	reg_lambda	1.0
11	min_child_samples	20
12	random_state	42
13	n_jobs	-1

Table 3. LightGBM Base Model Accuracy

Fold	Accuracy (%)	Precision	Recall	F1 Score	AUC
5	82.70	90.63	76.49	82.96	81.20
10	88.94	90.74	88.99	89.85	92.52

Model demonstrates reasonably good performance under both cross-validation scenarios, namely 5-fold and 10-fold. In the 5-fold cross-validation setting, the model achieves an accuracy of 82.70%, with a precision of 90.63% and a recall of 76.49%. This indicates that the model relatively rarely produces false positive predictions but still misses a portion of true positive cases (false negatives). This behavior is reflected in the F1 score of 82.96% and an AUC of 81.20%, which together suggest that the discriminative ability of the model falls within a good performance range.

In contrast, when the number of folds is increased to 10, there is a notable improvement across almost all evaluation metrics. Accuracy increases to 88.94%, precision remains stable at around 90.74%, while recall rises substantially to 88.99%. This improvement in recall indicates that the model becomes more capable of detecting patients with liver disease without sacrificing precision. The F1-score also increases to 89.85%, and the AUC reaches 92.52%, reflecting a forceful ability to separate positive and negative classes. In the context of disease detection, particularly liver disease, the 10-fold cross-validation configuration is therefore more advantageous, as it yields higher recall and AUC, thereby reducing the risk of failing to identify patients who actually have the disease.

To improve the performance of the model in diagnosing liver disease, researchers also use Optuna and Bayesian Optimizer as a hyperparameter in Tables 4 and 5, with cross validation of 5 and 10 so that the accuracy results were obtained as in Table 5.

Based on the experiments using the LightGBM model with the parameters shown in Table 3, Table 6 shows that Optuna yields a consistent improvement in performance under both 5-fold and 10-fold cross-validation settings. In the 5-fold scenario, accuracy increases to 83.88%, with a precision of 92.05%, recall of 77.45%, F1-score of 84.10%, and AUC of 81.58%, indicating that the model becomes slightly more sensitive to positive cases without sacrificing the correctness of its positive predictions. When the number of folds is increased to 10, the performance improves more substantially: accuracy rises to 90.76%, precision to 92.29%, recall to 90.81%, F1-score to 91.54%, and AUC to 93.03%.

Table 4. Parameters with Optuna

No	Parameter	Value
1	n_estimators	100 – 500
2	learning_rate	0.01 – 0.3
3	num_leaves	(15, 31, 63, 127)
4	max_depth	(-1, 3, 5, 7, 9)
5	subsample	0.6 – 1.0
6	colsample_bytree	0.6 – 1.0
7	reg_lambda	0.01 – 10.0
8	min_child_samples	5 – 50

Table 5. Parameters with Bayesian Optimization

No	Parameter	Value
1	n_estimators	100 – 500
2	learning_rate	0.01 – 0.3
3	num_leaves	15 – 127
4	max_depth	-1 – 9
5	subsample	0.6 – 1.0
6	colsample_bytree	0.6 – 1.0
7	reg_lambda	0.01 – 10.0
8	min_child_samples	5 – 50

Table 6. LightGBM Accuracy with Hyperparameters

Fold	Hyperparameter	Accuracy (%)	Precision	Recall	F1 Score	AUC
5	Optuna	83.88	92.05	77.45	84.10	81.58
10	Optuna	90.76	92.29	90.81	91.54	93.03
5	Bayesian	86.05	94.40	79.38	86.24	81.84
10	Bayesian	93.17	94.49	92.52	93.72	93.28

In contrast, Bayesian optimization produces results that are slightly superior to Optuna across all scenarios. Under 5-fold cross-validation, accuracy further increases to 86.05%, with very high precision (94.40%), recall of 79.38%, F1-score of 86.24%, and AUC of 81.84%. This indicates that the LightGBM model tuned with Bayesian optimization not only makes fewer false positive predictions for healthy patients (as reflected by the high precision) but is also more effective at identifying true positive cases compared with both the

baseline and the Optuna-tuned models in the 5-fold setting. Under 10-fold cross-validation, the model achieves its overall best performance: accuracy of 93.17%, precision of 94.49%, recall of 92.52%, F1-score of 93.72%, and AUC of 93.28%.

In general, the findings confirm that LightGBM performance depends considerably on the selection of its hyperparameter values. Applying Optuna and Bayesian optimization enables the model to identify more effective configurations, resulting in noticeably better predictive performance than that achieved using the default baseline parameters [20]. Nevertheless, this study still relies on a dataset of relatively limited size, which constrains the model's potential generalizability to broader populations. Future work employing larger datasets, combined with more systematic feature selection techniques and deeper hyperparameter exploration, is expected to produce liver disease detection models that are not only more accurate but also more stable and reliable to support clinical decision-making.

3.3 Discussion

The discussion of the experimental results indicates that the use of the LightGBM model for liver disease detection, when combined with a systematic hyperparameter optimization strategy, is capable of producing highly competitive classification performance. In general, the baseline model with a controlled default configuration already achieves satisfactory results; however, the experimental findings demonstrate that hyperparameter tuning using Optuna and Bayesian Optimization consistently enhances predictive quality in both the 5-fold and 10-fold cross-validation scenarios [19].

Both approaches are model-based search methods that optimize an objective function (in this case, AUC/ROC). Optuna, through its Tree-structured Parzen Estimator (TPE), models the distribution of hyperparameters associated with high versus low scores and selects candidate configurations with high Expected Improvement. This yields an adaptive search process that rapidly focuses on promising regions of the hyperparameter space (exploitation) while still probing new regions (exploration).

In contrast, Bayesian optimization approximates the objective function using a surrogate model and uses this approximation to select promising hyperparameter configurations. Consequently, it can approach an optimal solution with substantially fewer evaluations than exhaustive grid search or unguided random search. This mechanism may explain the consistent improvement observed across the evaluation metrics, particularly in the 10-fold cross-validation scenario. Rather than exploring hyperparameter combinations indiscriminately, Bayesian optimization incorporates information from earlier trials to determine which configurations should be evaluated next, thereby making the search process more efficient and directed.[20].

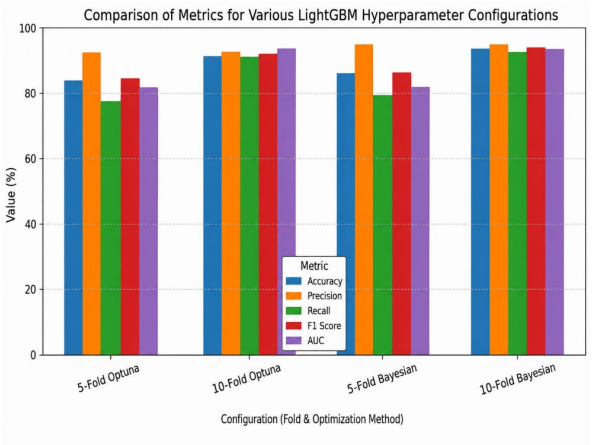


Figure 4. Summary of LightGBM Hyperparameter Configuration Results

Figure 4 compares accuracy, precision, recall, F1-score, and AUC across the Optuna- and Bayesian-optimized LightGBM configurations under 5-fold and 10-fold cross-validation. The chart shows that the 10-fold setting generally improves the metrics for both optimization methods, while the 10-fold Bayesian configuration produces the strongest and most balanced performance, with accuracy of 93.17%, precision of 94.49%, recall of 92.52%, F1-score of 93.72%, and AUC of 93.28%. The application of hyperparameter optimization using Optuna yielded a meaningful improvement in model performance. Under the 5-fold cross-validation setting, accuracy increased from 82.70% to 83.88%, precision from 90.63% to 92.05%, and recall from 76.49% to 77.45%, accompanied by an increase in F1-score to 84.10% and a slight improvement in AUC to 81.58%. The impact of Optuna becomes even more pronounced under 10-fold cross-validation, where accuracy rises to 90.76%, precision to 92.29%, recall to 90.81%, F1-score to 91.54%, and AUC to 93.03%. Compared with the 10-fold baseline, nearly all metrics show consistent gains, indicating that Optunathis approachable to identify a LightGBM configuration that achieves a better trade-off between model complexity and generalization ability, allowing the model to learn patterns that are truly relevant for liver disease diagnosis rather than merely memorizing the training data.

Bayesian optimization provides an even more substantial performance boost compared with Optuna, particularly when considering the combination of metrics most critical for disease detection. Under 5-fold cross-validation, accuracy increases to 86.05%, with very high precision (94.40%), recall of 79.38%, F1-score of 86.24%, and AUC of 81.84%. The high precision indicates that almost all patients predicted as having liver disease are truly positive, thereby minimizing the risk of incorrectly labeling healthy patients as sick. At the same time, recall improves relative to both the baseline and Optuna-tuned models, meaning that more true cases of liver disease are

successfully detected. The best overall performance is achieved with Bayesian Optimization under 10-fold cross-validation, yielding an accuracy of 93.17%, precision of 94.49%, recall of 92.52%, F1-score of 93.72%, and AUC of 93.28%.

This configuration represents the most balanced and robust result, combining high discriminative ability with low classification error for both positive and negative cases. Additionally, as shown in Table 7, our optimized LightGBM models are more accurate than previous machine learning studies on Non-Alcoholic Fatty Liver Disease (NAFLD) datasets. This shows that our approach is competitive and could help doctors make better decisions about liver disease.

Table 7. Compare Our Work Results with the Previous Results

Title	ML Algorithm	Dataset	Best Acc	Year
Attention-driven graph-based machine learning for non-invasive diagnosis of NAFLD	KNN	NAFLD	75.02%	2025
Large-Scale Identification and Classification of Non-Alcoholic Fatty Liver Disease Using Machine Learning	LR, RF, Xgboost, Nb	NAFLD	88.80%	2022
The Effect of Data Balancing Techniques on NAFLD Classification Using the SVM Algorithm	SVM, SMOTE	NAFLD	78.70%	2025
Machine Learning-Based Prediction of Non-Alcoholic Fatty Liver Disease Using NHANES Data	Ensemble boosted trees	NAFLD	71.01%	2021
Our Study	LightGBM,Hyperparameter	NAFLD	93.01%	

Based on the comparative summary in the table, the proposed model in this study, LightGBM with hyperparameter optimization, achieves an accuracy of 93.01% for NAFLD detection, outperforming all prior works used as benchmarks. The study in [21] which employed K-Nearest Neighbors (KNN) for non-invasive NAFLD diagnosis, reported an accuracy of only 75.02%, indicating that neighbor-based methods tend to struggle in capturing complex, non-linear patterns in clinical data, especially when the number of features and inter-patient variability are relatively high.

The work in [22], which utilized Support Vector Machines (SVM) combined with SMOTE for data balancing, achieved a slightly better accuracy of 78.70%, yet still lags significantly behind the performance of the proposed model. This gap may be attributed to inherent limitations of SVM in modeling intricate feature interactions without extensive feature engineering, as well as its sensitivity to kernel choice and parameter settings, which can constrain its effectiveness compared to modern gradient boosting approaches such as LightGBM. The study in [23] which employed ensemble boosted trees on the NHANES database, reported an even lower accuracy of 71.01%. Although this approach also relies on tree-based ensembles, its relatively modest performance may be influenced by differences in population characteristics, predictor variables, and evaluation design, but it also suggests that simply using boosted trees without thorough hyperparameter tuning is not sufficient to achieve optimal results. Meanwhile, the work in [24], which combined several classical algorithms such as Logistic Regression, Random Forest, XGBoost, and Naive Bayes, obtained a higher accuracy of 88.80% and can be regarded as one of the stronger baselines in the NAFLD literature.

Nevertheless, this value still falls below the accuracy achieved by the LightGBM-based model in the present study. These findings indicate that even when prior studies have utilized relatively powerful algorithms (e.g., RF and XGBoost), there remains considerable room for improvement through the selection of a more efficient gradient boosting framework and the application of more systematic hyperparameter tuning strategies.

The strength of the proposed model in this study lies not only in the choice of LightGBM as a more modern and efficient variant of gradient boosting, but also in the use of targeted hyperparameter optimization strategies (Optuna and Bayesian Optimization) and a rigorous cross-validation scheme (5- and 10-fold). LightGBM is specifically designed to handle tabular data with mixed feature types and complex non-linear relationships, while hyperparameter optimization enables the discovery of configurations (e.g., number of trees, tree depth, learning rate, and regularization strength) that are well aligned with the characteristics of the NAFLD dataset used in this work [25].

This is reflected in the consistent improvement in accuracy across different fold settings, as well as in the high precision, recall, and F1-score, indicating that the model is not only accurate in aggregate but also well balanced between detecting NAFLD cases and avoiding misclassification of healthy patients. Thus, although differences in population, feature sets, and evaluation protocols across studies mean that accuracy figures are not fully comparable in an “apple-to-apple” manner, the overall pattern clearly suggests that the LightGBM approach with hybrid hyperparameter optimization proposed in this study delivers a tangible performance gain over several comparative methods previously reported in the NAFLD literature.

4. Conclusions

Based on the series of experiments conducted, this study concludes that the LightGBM model with hybrid hyperparameter optimization using Optuna and Bayesian Optimization can significantly improve the performance of liver disease detection compared to the baseline configuration. By utilizing 10 clinical and lifestyle features from 1,700 patient records and an Out-of-Fold evaluation scheme based on Stratified Shuffle

Split (5-fold and 10-fold), the baseline model already demonstrated reasonably good performance, yet its results could still be enhanced. The application of Optuna improved almost all evaluation metrics, while Bayesian Optimization yielded the best overall gains, particularly in the 10-fold cross-validation setting, achieving an accuracy of around 93%, precision and recall above 92%, F1-score above 93%, and AUC above 93%.

The obtained results demonstrate a favorable balance between the model's ability to detect individuals with liver disease and its capacity to avoid incorrectly classifying healthy subjects. These findings answer the primary research question by confirming that systematic hyperparameter tuning is an important factor in improving the classification performance of LightGBM. The results also suggest that the proposed method achieves better predictive performance than several models previously evaluated for comparable liver disease classification tasks.

From a methodological perspective, the principal contribution of this study is the combination of LightGBM with a hybrid hyperparameter optimization framework involving Optuna and Bayesian Optimization. This framework is further supported by a rigorous out-of-fold evaluation procedure, which produces a more stable performance estimate and reduces the model's dependence on a particular train–test split.

From a practical standpoint, the developed model may serve as a supporting instrument for the early identification of liver disease risk. By analyzing relevant clinical and lifestyle-related variables, the model could assist healthcare professionals in identifying patients who may require further examination. However, its predictions should be regarded as supplementary information rather than a substitute for clinical diagnosis.

Several limitations should also be acknowledged. The experiments were conducted using only one dataset of moderate size, and the evaluation was restricted to internal validation. Future studies should therefore examine the proposed method using larger and more heterogeneous datasets, conduct external validation involving multiple healthcare institutions, and investigate feature-selection and model-interpretability methods such as SHAP. Further evaluation within actual clinical environments is also required to determine the feasibility, reliability, and practical value of integrating the model into routine healthcare workflows.

References

[1] L. Yao *et al.*, "Global trends of big data analytics in health research: a bibliometric study," *Front. Med. (Lausanne).*, vol. 12, Jul. 2025, doi: https://doi.org/10.3389/fmed.2025.1456286 .

[2] N. Shibata *et al.*, "A machine learning algorithm for stratification of risk of cardiovascular disease in metabolic dysfunction-associated steatotic liver disease," *Eur. J. Intern. Med.*, vol. 129, pp. 62–70, Nov. 2024, doi: https://doi.org/10.1016/j.ejim.2024.07.005.

[3] G. Cao, J. Liu, and M. Liu, "Global, Regional, and National Trends in Incidence and Mortality of Primary Liver Cancer and Its Underlying Etiologies from 1990 to 2019: Results from the Global Burden of Disease Study 2019," *J. Epidemiol. Glob. Health*, vol. 13, no. 2, pp. 344–360, May 2023, doi: https://doi.org/10.1007/s44197-023-00109-0.

[4] H. Zhang *et al.*, "Development and validation of machine learning-augmented algorithm for insulin sensitivity assessment in the community and primary care settings: a population-based study in China," *Front. Endocrinol. (Lausanne).*, vol. 15, Jan. 2024, doi: https://doi.org/10.3389/fendo.2024.1292346.

[5] S. Hanifi, A. Cammarono, and H. Zare-Behtash, "Advanced hyperparameter optimization of deep learning models for wind power prediction," *Renew. Energy*, vol. 221, p. 119700, Feb. 2024, doi: https://doi.org/10.1016/j.renene.2023.119700.

[6] P. Kumar Sahu and T. Fatma, "Optimized Breast Cancer Classification Using PCA-LASSO Feature Selection and Ensemble Learning Strategies With Optuna Optimization," *IEEE Access*, vol. 13, pp. 35645–35661, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3539746.

[7] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, "Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification," *Sensors*, vol. 23, no. 4, p. 2085, Feb. 2023, doi: https://doi.org/10.3390/s23042085.

[8] H. Kim and D. Sanz-Alonso, "Enhancing Gaussian Process Surrogates for Optimization and Posterior Approximation via Random Exploration," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 13, no. 3, pp. 1054–1084, Sep. 2025, doi: https://doi.org/10.1137/24M1677009.

[9] S. Agrawal and D. K. Sharma, "Handling High-Dimensional Data and Classification Using a Hybrid Feature Selection Approach," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, Apr. 2022, pp. 2168–2172. doi: https://doi.org/10.1109/ICACITE53722.2022.9823878.

[10] H. Wu, Y. Mao, J. Weng, Y. Yu, and J. Wang, "Fractional light gradient boosting machine ensemble learning model: A non-causal fractional difference descent approach," *Information Fusion*, vol. 118, p. 102947, Jun. 2025, doi: https://doi.org/10.1016/j.inffus.2025.102947.

[11] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A Novel Scheme for Mapping of MVT-Type Pb–Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm," *Natural Resources Research*, vol. 32, no. 6, pp. 2417–2438, Dec. 2023, doi: https://doi.org/10.1007/s11053-023-10249-6.

[12] G. He, Z. Wang, L. Tang, W. Yu, F. Nie, and X. Li, "Reweighted-Boosting: A Gradient-Based Boosting Optimization Framework," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 11953–11965, Jul. 2025, doi: https://doi.org/10.1109/TNNLS.2024.3457764 .

[13] F. Yamamoto, S. Ozawa, and L. Wang, "eFL-Boost: Efficient Federated Learning for Gradient Boosting Decision Trees," *IEEE Access*, vol. 10, pp. 43954–43963, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3169502.

[14] I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim, "An efficient churn prediction model using gradient boosting machine and metaheuristic optimization," *Sci. Rep.*, vol. 13, no. 1, p. 14441, Sep. 2023, doi: https://doi.org/10.1038/s41598-023-41093-6.

[15] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, Apr. 2023, doi: https://doi.org/10.1002/eng2.12599.

[16] S. Hanifi, A. Cammarono, and H. Zare-Behtash, "Advanced hyperparameter optimization of deep learning models for wind power prediction," *Renew. Energy*, vol. 221, p. 119700, Feb. 2024, doi: https://doi.org/10.1016/j.renene.2023.119700.

[17] P. R, S. Obbayed, K. Manju, G. Swarnalakshmi, and N. Purushotham, "Employee Attrition Prediction based on Light Gradient Boosting Machine with Bayesian Optimization," in *2025 International Conference on Intelligent Systems and*

Computational Networks (ICISCN), IEEE, Jan. 2025, pp. 1–5. doi: https://doi.org/10.1109/ICISCN64258.2025.10934683.

[18] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: https://doi.org/10.3390/s23042333.

[19] U. N. Yefou, P. O. M. Choudja, B. Sow, and A. Adejumo, "Optimized Machine Learning Models for Hepatitis C Prediction: Leveraging Optuna for Hyperparameter Tuning and Streamlit for Model Deployment," 2024, pp. 88–100. doi: https://doi.org/10.1007/978-3-031-57624-9_5.

[20] D. T. Anh *et al.*, "Assessment of groundwater potential modeling using support vector machine optimization based on Bayesian multi-objective hyperparameter algorithm," *Appl. Soft Comput.*, vol. 132, p. 109848, Jan. 2023, doi: https://doi.org/10.1016/j.asoc.2022.109848.

[21] E. Srivastava, S. Mohan, T. K. Gandhi, A. K. Choudhury, and S. Kumar, "Attention-driven graph-based machine learning for non-invasive diagnosis of NAFLD," *Intell. Based. Med.*, vol. 12, Jan. 2025, doi: https://doi.org/10.1016/j.ibmed.2025.100288.

[22] R. Mahardika Faska, S. Kurnia Gusti, E. Budianita, and F. Syafria, "PENGARUH TEKNIK PENYEIMBANGAN DATA PADA KLASIFIKASI PENYAKIT NAFLD DENGAN ALGORITMA SVM," *Jurnal Informatika Teknologi dan Sains*, vol. 7, no. 2, pp. 858–867, 2025, [Online]. Available: https://www.kaggle.com/datasets/utkarshx27/non-alcohol-

[23] A. Atsawarungruangkit, P. Laoveeravat, and K. Promrat, "Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database," *World J. Hepatol.*, vol. 13, no. 10, pp. 1418–1427, Oct. 2021, doi: https://doi.org/10.4254/wjh.v13.i10.1417.

[24] W. Ji, M. Xue, Y. Zhang, H. Yao, and Y. Wang, "A Machine Learning Based Framework to Identify and Classify Non-alcoholic Fatty Liver Disease in a Large-Scale Population," *Front. Public Health*, vol. 10, Apr. 2022, doi: https://doi.org/10.3389/fpubh.2022.846118.

[25] C. Chen and H. Seo, "Prediction of rock mass class ahead of TBM excavation face by ML and DL algorithms with Bayesian TPE optimization and SHAP feature analysis," *Acta Geotech.*, vol. 18, no. 7, pp. 3825–3848, Jul. 2023, doi: https://doi.org/10.1007/s11440-022-01779-z.