Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 11, 2026: 4663-4676
ISSN:2302-8149
e-ISSN:2540-9719

Perbandingan *MLP, AdaBoost,* dan *CatBoost* dengan *SMOTE–ENN* untuk Prediksi Penyakit Hati

Comparison of MLP, AdaBoost, and CatBoost with SMOTE–ENN for Liver Disease Prediction

[1]Muhammad Satrio Kurniawan, [2]Muhamad Akrom*
[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Semarang, Jawa Tengah, Indonesia
*e-mail: m.akrom@dsn.dinus.ac.id

(***received*****:** 18 October 2026, ***revised*****:** 22 October 2026, ***accepted*****:** 24 October 2026)

Abstrak

Penyakit hati merupakan salah satu permasalahan kesehatan global dengan tingkat prevalensi dan mortalitas yang tinggi, sehingga diagnosis dini dan akurat menjadi sangat penting untuk meningkatkan hasil klinis dan menurunkan angka kematian. Penelitian ini mengusulkan kerangka prediksi penyakit hati berbasis *machine learning* menggunakan data klinis dari Indian Liver Patient Dataset (ILPD), yang memiliki karakteristik ketidakseimbangan kelas. Untuk mengatasi permasalahan tersebut, diterapkan teknik *hybrid resampling* Synthetic Minority Oversampling Technique dan Edited Nearest Neighbors (SMOTE–ENN) pada data pelatihan. Penelitian ini mengevaluasi dan membandingkan kinerja tiga model *machine learning*, yaitu Multilayer Perceptron (MLP), AdaBoost, dan CatBoost. Proses optimasi hiperparameter dilakukan menggunakan GridSearchCV dengan skema *stratified cross-validation*. Kinerja model dievaluasi menggunakan beberapa metrik, meliputi accuracy, precision, recall, F1-score, dan receiver operating characteristic area under the curve (ROC–AUC), dengan ROC–AUC digunakan sebagai kriteria utama dalam pemilihan model terbaik. Hasil eksperimen menunjukkan bahwa model MLP memberikan kinerja terbaik secara keseluruhan dengan akurasi sebesar 75% dan nilai ROC–AUC tertinggi sebesar 0,848, yang mengindikasikan kemampuan diskriminasi kelas yang lebih baik dibandingkan dengan model AdaBoost dan CatBoost. Temuan ini menunjukkan bahwa pendekatan yang diusulkan efektif dalam meningkatkan performa prediksi penyakit hati pada data klinis yang tidak seimbang, serta menegaskan pentingnya penggunaan metrik ROC–AUC dalam sistem pendukung keputusan medis.

Kata kunci: prediksi penyakit hati, *machine learning*, SMOTE–ENN, ILPD, ROC–AUC

Abstract

Liver disease is a major global health concern, characterized by high prevalence and mortality rates. Therefore, early and accurate diagnosis is essential for improving clinical outcomes and reducing mortality. This study proposes a machine learning-based framework for liver disease prediction using clinical data from the Indian Liver Patient Dataset (ILPD), which exhibits class imbalance. To address this issue, a hybrid resampling technique combining the Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors (SMOTE–ENN) was applied to the training data. This study evaluates and compares the performance of three machine learning models: Multilayer Perceptron (MLP), AdaBoost, and CatBoost. Hyperparameter optimization was performed using GridSearchCV with a stratified cross-validation strategy. Model performance was evaluated using several metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC–AUC). ROC–AUC was used as the primary criterion for selecting the best-performing model. The experimental results demonstrate that MLP achieved the best overall performance, with an accuracy of 75% and the highest ROC–AUC value of 0.848, indicating better class discrimination than AdaBoost and CatBoost. These findings suggest that the proposed approach has the potential to improve liver disease prediction performance on imbalanced clinical datasets and highlight the importance of using ROC–AUC as an evaluation metric in medical decision support systems.

Keywords: *liver disease prediction, machine learning, SMOTE–ENN, ILPD, ROC–AUC*

http://sistemasi.ftik.unisi.ac.id

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Hati merupakan organ vital yang berperan penting dalam proses metabolisme, penyerapan nutrisi, dan detoksifikasi zat-zat berbahaya yang masuk ke dalam tubuh, sehingga penurunan fungsi kerja hati dapat menyebabkan terjadinya penyakit hati yang serius. Penyakit hati disebabkan oleh berbagai faktor, seperti faktor genetik, pengaruh lingkungan, serta kebiasaan hidup yang tidak sehat, dan mencakup berbagai kondisi seperti hepatitis (A, B, C, D, dan E), penyakit lemak hati, sirosis, kanker hati, kegagalan fungsi hati, penyakit Wilson, hemochromatosis, dan defisiensi alpha-1 antitripsin [1]. Penyakit hati tercatat sebagai masalah kesehatan global, dengan lebih dari satu juta kasus baru yang didiagnosis setiap tahun, dan kondisi ini dapat menyebabkan kematian, terutama di negara dengan pendapatan rendah hingga menengah [2]. Pada tahun 2017, sirosis hati sendiri diperkirakan menyebabkan hampir mendekati 1,32 juta kematian di seluruh dunia, yang merepresentasikan sekitar 3,5% dari angka kematian global, dengan prevalensi yang lebih tinggi diamati di kalangan pria [3]. Meskipun terdapat peningkatan fasilitas medis dan penerapan gaya hidup yang lebih sehat yang berpengaruh terhadap penurunan angka kematian secara bertahap, kematian yang berhubungan dengan penyakit hati masih tetap tinggi, yang sebagian besar disebabkan oleh faktor risiko seperti konsumsi alkohol berlebihan, penggunaan narkoba berkepanjangan, obesitas, dan diabetes [4]. Oleh karena itu, diagnosis dini dan akurat pada penyakit hati sangat penting untuk mengurangi risiko penyakit, meningkatkan hasil klinis, dan menurunkan angka kematian.

Seiring berkembangnya kecerdasan buatan, khususnya *machine learning* (ML), pemanfaatan data klinis yang mencakup rekam medis elektronik, hasil laboratorium, citra medis, hingga sinyal fisiologis dapat diolah menjadi model prediktif yang membantu tenaga kesehatan dalam mendeteksi penyakit lebih dini, mengestimasi risiko komplikasi, serta mendukung pengambilan keputusan secara lebih objektif dan konsisten [5], [6]. Dalam kerangka ini, ML mampu mempelajari pola-pola kompleks yang sering kali sulit ditangkap melalui penilaian manual semata, sehingga berpotensi meningkatkan akurasi diagnosis dan mempercepat proses triase maupun tindak lanjut pasien [7]. Selain itu, pendekatan ML memungkinkan analisis data secara otomatis dan berkelanjutan, sehingga menjadikannya sangat relevan untuk penanganan penyakit kronis dan multifaktorial yang memerlukan pemantauan jangka panjang [8]. Integrasi ML sebagai *clinical decision support system* juga dipandang dapat meningkatkan efisiensi layanan kesehatan tanpa menggantikan peran klinisi, melainkan sebagai alat bantu dalam pengambilan keputusan berbasis data [9].

Penelitian sebelumnya telah banyak mengeksplorasi penerapan metode *machine learning* untuk prediksi penyakit hati menggunakan data klinis, khususnya *Indian Liver Patient Dataset* (ILPD). Gupta et al. [10] memanfaatkan dataset ILPD dari *UCI Machine Learning Repository* dengan menerapkan berbagai algoritma *supervised learning*, seperti Logistic Regression, Decision Tree, Random Forest, KNN, Gradient Boosting, XGBoost, LightGBM, serta metode *stacking*, dan melaporkan akurasi tertinggi sebesar 63% yang dicapai oleh Random Forest dan LightGBM. Selanjutnya, Altaf et al. [11] menggunakan dataset ILPD dan PIMA Indians Diabetes dengan pendekatan *voting ensemble* yang mengombinasikan XGBoost, Bagging Meta-Estimator, dan Multilayer Perceptron (MLP), serta menerapkan metode seleksi fitur *Mean Decrease in Impurity* (MDI), sehingga memperoleh peningkatan akurasi hingga 73,56% pada dataset ILPD. Sementara itu, Hikmah et al. [12] mengevaluasi sejumlah algoritma klasifikasi, termasuk Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, k-NN, dan Logistic Regression, menggunakan perangkat lunak *Orange Data Mining*, dengan Logistic Regression sebagai model terbaik dengan akurasi 71,8%. Meskipun berbagai pendekatan tersebut telah menunjukkan potensi *machine learning* dalam prediksi penyakit hati, sebagian besar penelitian masih berfokus pada metrik akurasi dan belum secara komprehensif menangani permasalahan ketidakseimbangan kelas serta keberadaan *noise* dan *overlap* antarkelas yang menjadi karakteristik utama dataset ILPD.

Berdasarkan latar belakang tersebut, pada penelitian ini dikembangkan sebuah kerangka eksperimen yang menekankan penanganan ketidakseimbangan data melalui penerapan teknik hybrid resampling SMOTE–ENN pada data pelatihan. Penelitian ini membandingkan kinerja beberapa model *machine learning*, yaitu Multilayer Perceptron (MLP), AdaBoost, dan CatBoost, serta menerapkan optimasi hiperparameter menggunakan metode GridSearchCV yang dikombinasikan dengan skema cross-validation. Kinerja model dievaluasi secara menyeluruh menggunakan metrik accuracy, precision, recall, F1-score, serta ROC–AUC untuk memberikan analisis yang lebih objektif dan

 ISSN:2302-8149
e-ISSN:2540-9719

komprehensif terhadap kemampuan klasifikasi dan diskriminasi model pada dataset yang tidak seimbang.

2 Tinjauan Literatur

Penelitian prediksi penyakit hati menggunakan *Indian Liver Patient Dataset* (ILPD) telah mengeksplorasi berbagai algoritma *machine learning* dan strategi pengolahan data. Gupta et al. [10] menerapkan imputasi median, eliminasi *outlier*, SMOTE, dan seleksi fitur berbasis Random Forest pada sepuluh algoritma, termasuk *stacking*. Namun, akurasi yang diperoleh hanya berkisar antara 54% hingga 63%, dengan Random Forest dan LightGBM sebagai model terbaik. Hasil serupa dalam hal variasi kinerja model ditemukan oleh Hikmah et al. [12], yang membandingkan tujuh algoritma menggunakan Orange dengan pembagian data 80:20 yang diulang sebanyak 100 kali. Logistic Regression menghasilkan akurasi tertinggi sebesar 71,8%, sedikit di atas Neural Network sebesar 71,1% dan Random Forest sebesar 70,2%. Sementara itu, Altaf et al. [11] menggabungkan XGBoost, Bagging, dan Multilayer Perceptron (MLP) dalam pendekatan *hard voting ensemble*. Penerapan seleksi fitur *Mean Decrease in Impurity* meningkatkan akurasi model dari 65,52% menjadi 73,56%, sedangkan akurasi MLP meningkat dari 59,20% menjadi 70,11%. Temuan-temuan tersebut memperlihatkan bahwa penggunaan algoritma yang lebih kompleks tidak selalu menghasilkan kinerja yang lebih baik. Strategi pengolahan data juga berperan penting, terutama karena ILPD memiliki distribusi kelas yang tidak seimbang, yaitu 416 data pasien penyakit hati dan 167 data pasien non-penyakit hati.

Meskipun beberapa penelitian melaporkan akurasi yang relatif baik, kinerja model dalam mengenali kelas minoritas masih menjadi tantangan. Hikmah et al. [12] melaporkan bahwa Logistic Regression dan Random Forest hanya mengklasifikasikan dengan benar sekitar 24,6% dan 14,9% kasus non-penyakit hati. Sementara itu, Altaf et al. [11] memperoleh *recall* sebesar 0,28 untuk kelas non-penyakit hati, jauh di bawah kelas penyakit hati sebesar 0,92, serta menunjukkan indikasi *overfitting*. Kedua penelitian tersebut belum menerapkan penanganan ketidakseimbangan kelas secara khusus. Gupta et al. [10] telah menggunakan SMOTE, tetapi prosedur penerapannya dan evaluasi kinerja per kelas belum dijelaskan secara memadai. Selain itu, seleksi fitur sebelum pembagian data pada Altaf et al. [11] berpotensi menimbulkan *data leakage*, sementara optimasi hiperparameter secara sistematis belum dilaporkan. Temuan ini menunjukkan perlunya strategi penyeimbangan data dan evaluasi model yang lebih cermat untuk menghasilkan prediksi yang lebih andal.

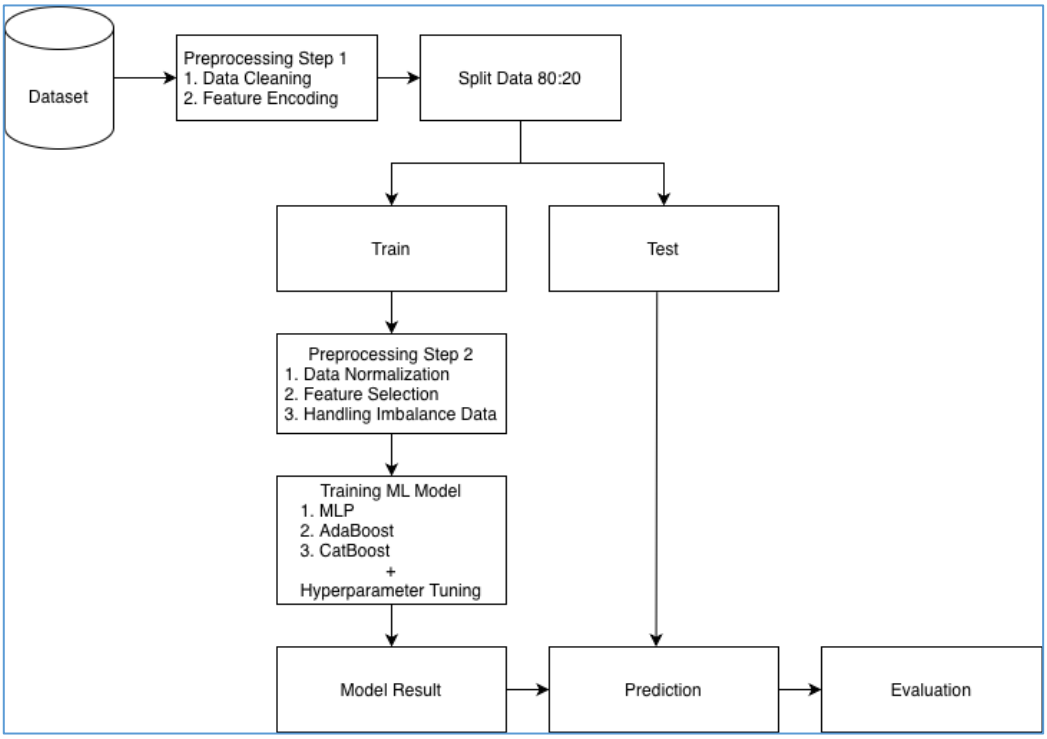
Salah satu pendekatan yang dapat dipertimbangkan untuk mengatasi ketidakseimbangan kelas sekaligus mengurangi sampel yang berpotensi mengganggu proses pembelajaran adalah *hybrid resampling* SMOTE–ENN. Metode ini menggabungkan Synthetic Minority Oversampling Technique (SMOTE), yang membentuk sampel sintetis pada kelas minoritas, dengan Edited Nearest Neighbors (ENN), yang menghapus sampel berdasarkan ketidaksesuaian kelas dengan tetangga terdekatnya [13]. Rani et al. [14] telah mengevaluasi SMOTE–ENN pada ILPD dan melaporkan hasil yang lebih baik dibandingkan beberapa teknik penyeimbangan lainnya. Namun, penyeimbangan data dilakukan sebelum pembagian data pelatihan dan pengujian, sehingga berpotensi menyebabkan *data leakage* dan mengubah distribusi kelas pada data pengujian. Berdasarkan keterbatasan tersebut, penelitian ini menerapkan SMOTE–ENN hanya pada data pelatihan agar evaluasi tetap dilakukan pada data pengujian dengan distribusi asli. Penelitian ini juga membandingkan MLP, AdaBoost, dan CatBoost yang dioptimasi menggunakan GridSearchCV dengan *stratified cross-validation*. Kinerja model dinilai berdasarkan *accuracy*, *precision*, *recall*, F1-score, dan *receiver operating characteristic area under the curve* (ROC–AUC), dengan ROC–AUC sebagai kriteria utama pemilihan model terbaik. Dengan demikian, penelitian ini berfokus pada evaluasi kombinasi *hybrid resampling*, optimasi hiperparameter, dan perbandingan model untuk mempelajari pengaruhnya terhadap prediksi penyakit hati pada ILPD. Hipotesis yang diajukan adalah bahwa penerapan SMOTE–ENN pada data pelatihan dapat meningkatkan *recall* dan F1-score kelas minoritas dibandingkan model yang dilatih tanpa penanganan ketidakseimbangan kelas.

3 Metode Penelitian

Alur kerja keseluruhan dari metodologi yang diusulkan dalam penelitian ini ditunjukkan pada Gambar 1. Proses diawali dengan pengumpulan dataset, kemudian dilanjutkan dengan Preprocessing

 ISSN:2302-8149
e-ISSN:2540-9719

Step 1 yang mencakup *data cleaning* dan *feature encoding* untuk memastikan kualitas serta kompatibilitas data dengan algoritma *machine learning*. Dataset yang telah dipraproses selanjutnya dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20. Setelah itu, Preprocessing Step 2 diterapkan khusus pada data pelatihan, yang meliputi normalisasi data, seleksi fitur, dan penanganan ketidakseimbangan kelas guna mengoptimalkan proses pembelajaran model. Data pelatihan yang telah diproses digunakan untuk melatih beberapa model *machine learning*, yaitu MLP, AdaBoost, dan CatBoost, disertai dengan proses *hyperparameter tuning*. Tahap akhir meliputi proses prediksi menggunakan data pengujian serta evaluasi kinerja model menggunakan beberapa metrik evaluasi yaitu accuracy, precision, recall, F1-score, serta ROC–AUC.



Gambar 1 Tahapan penelitian

3.1 Dataset

Dataset Indian Liver Patient Dataset (ILPD) dikumpulkan dari wilayah Timur Laut Andhra Pradesh, India, dan terdiri atas 583 data observasi yang mencakup 10 fitur serta 1 variabel target. Dataset tersebut diperoleh dari UCI Machine Learning Repository sebagai sumber data utama dalam penelitian ini [15]. Variabel target pada dataset ILPD merepresentasikan status penyakit hati pasien, yang dikategorikan ke dalam dua kelas, yaitu pasien dengan penyakit hati (*liver*) dan pasien tanpa penyakit hati (*non-liver*). Dari total 583 pasien, 416 terdiagnosis penyakit hati dan 167 tidak terdiagnosis penyakit hati. Distribusi demografis berdasarkan jenis kelamin menunjukkan 441 pasien laki-laki dan 142 pasien perempuan.

Fitur-fitur pada ILPD mencakup karakteristik demografis, klinis, dan biokimia pasien yang berkaitan dengan fungsi hati. Fitur tersebut meliputi Age (usia pasien) dan Gender (jenis kelamin pasien) sebagai faktor demografis, TB (kadar total bilirubin dalam darah) dan DB (kadar direct bilirubin dalam darah) sebagai indikator metabolisme bilirubin, Alkphos (kadar alkaline phosphatase), Sgpt (kadar alanine aminotransferase/ALT) dan Sgot (kadar aspartate aminotransferase/AST) yang mencerminkan tingkat kerusakan atau gangguan fungsi sel hati, serta TP (total protein dalam darah), ALB (kadar albumin dalam darah), dan A/G Ratio (rasio albumin terhadap globulin) yang menggambarkan kemampuan sintetis protein oleh hati dan keseimbangan protein serum darah.

 ISSN:2302-8149
e-ISSN:2540-9719

3.2 Data Splitting

Dataset dibagi menjadi data latih (*training set*) dan data uji (*testing set*) untuk mengevaluasi kinerja model secara objektif, dengan proporsi 80% data latih dan 20% data uji. Proses pembagian data dilakukan menggunakan strategi *stratified splitting* guna mempertahankan distribusi kelas asli pada kedua subset. Selain itu, seluruh proses pembagian data menggunakan nilai *random seed* tetap untuk memastikan reprodusibilitas hasil eksperimen. Pembagian data ini penting untuk mengurangi risiko overfitting, yaitu kondisi ketika model menunjukkan kinerja yang sangat baik pada data latih namun memiliki kemampuan generalisasi yang rendah terhadap data yang belum pernah dilihat sebelumnya [16].

3.3 Preprocessing

3.3.1 Preprocessing Step 1

Pada Preprocessing Step 1, dilakukan tahap pembersihan data dan pengolahan fitur kategorikal sebelum dataset dibagi menjadi data pelatihan dan data pengujian. Tahap ini bertujuan untuk memastikan kualitas data tanpa melibatkan pembelajaran parameter statistik yang dapat menyebabkan *data leakage*. Hasil pemeriksaan menunjukkan adanya 4 *missing value* dan 13 data duplikat, sehingga seluruh observasi tersebut dihapus dari dataset. Selanjutnya, fitur kategorikal *Gender* dilakukan proses *encoding* dengan mengonversi nilai "*Female*" menjadi 0 dan "*Male*" menjadi 1 agar kompatibel dengan algoritma *machine learning*. Tahap ini menghasilkan dataset yang bersih dan siap untuk proses pembagian data.

3.3.2 Preprocessing Step 2

Pada Preprocessing Step 2, dilakukan tahap praproses lanjutan setelah dataset dibagi menjadi data pelatihan dan data pengujian, dengan tujuan menyiapkan data agar optimal untuk proses pemodelan *machine learning*. Tahap pertama adalah penskalaan (*scaling*) fitur numerik menggunakan metode Robust Scaler. Robust Scaler melakukan penskalaan fitur dengan meminimalkan pengaruh nilai *outlier* melalui pendekatan statistik yang bersifat *robust*. Berbeda dengan Standard Scaler yang menggunakan rata-rata (μ) dan simpangan baku (σ), Robust Scaler memusatkan data pada median (m) dan melakukan penskalaan menggunakan *interquartile range* (IQR), yaitu selisih antara kuartil ketiga (Q_3) dan kuartil pertama (Q_1) [17]. Pendekatan ini membuat Robust Scaler lebih tahan terhadap distribusi data yang tidak normal serta keberadaan nilai ekstrem yang umum dijumpai pada dataset medis dan klinis. Persamaannya dapat dilihat pada Persamaan 1.

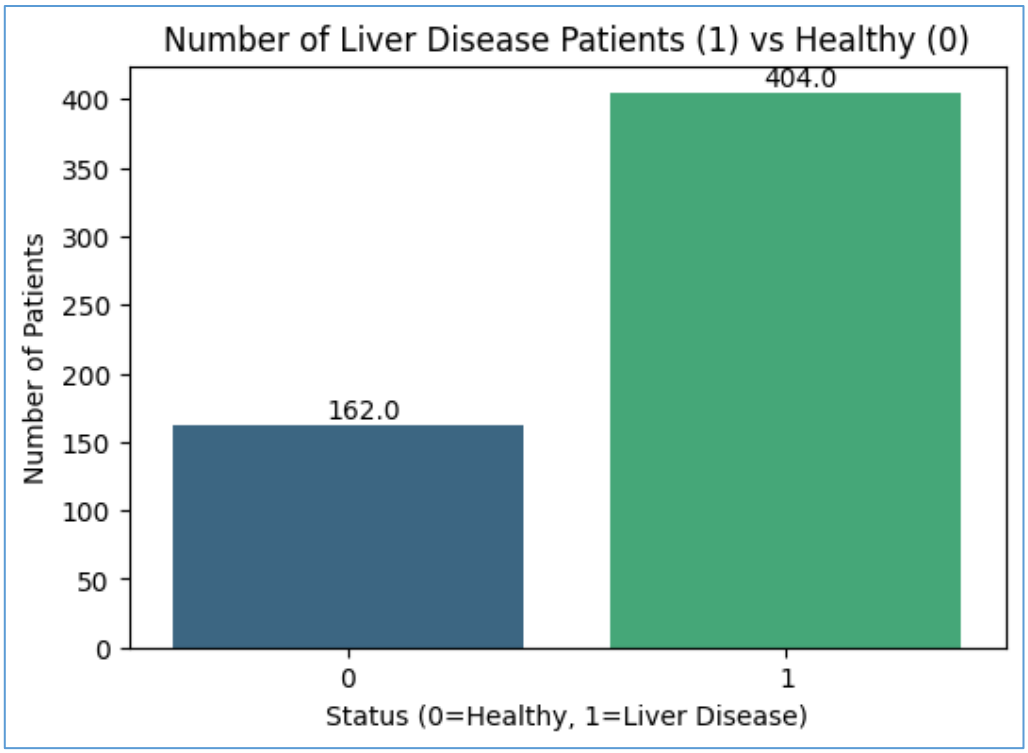
$$X_{scaled} = \frac{X - median(X)}{Q3 - Q1} \quad (1)$$

Tahap kedua adalah seleksi fitur menggunakan Recursive Feature Elimination (RFE) dengan estimator Random Forest. RFE [18] bekerja dengan cara melatih model secara iteratif dan mengeliminasi fitur dengan tingkat kepentingan (*feature importance*) terendah pada setiap iterasi hingga tersisa jumlah fitur yang telah ditentukan. Random Forest digunakan sebagai *base estimator* karena kemampuannya dalam menangkap hubungan nonlinier serta memberikan estimasi kepentingan fitur yang stabil [19]. Dalam implementasi ini, RFE dikonfigurasikan untuk memilih delapan fitur terbaik dari keseluruhan fitur masukan, dengan proses eliminasi dilakukan secara bertahap (*step* = 1). Seluruh proses seleksi fitur diterapkan hanya pada data pelatihan yang telah melalui tahap penskalaan untuk menghindari *data leakage*. Dari hasil penerapan RFE, fitur yang terpilih adalah adalah Age, TB, DB, Alkphos, Sgpt, Sgot, TP, dan ALB. Subset fitur ini kemudian digunakan pada tahap pelatihan dan evaluasi model klasifikasi selanjutnya guna meningkatkan kinerja prediksi, mengurangi kompleksitas model, dan meningkatkan interpretabilitas hasil.

Tahap ketiga adalah penanganan ketidakseimbangan kelas. Berdasarkan distribusi kelas yang ditunjukkan pada Gambar 2, dataset ILPD terdiri dari 404 kasus pasien dengan penyakit hati (kelas positif) dan 162 kasus individu sehat (kelas negatif). Untuk itu diterapkan metode SMOTE-ENN. SMOTE-ENN merupakan metode *hybrid resampling* yang mengombinasikan teknik *oversampling*

 ISSN:2302-8149
e-ISSN:2540-9719

Synthetic Minority Over-sampling Technique (SMOTE) dengan *undersampling* Edited Nearest Neighbors (ENN) untuk menangani permasalahan ketidakseimbangan kelas secara lebih efektif. Pada tahap pertama, SMOTE menghasilkan sampel sintetis pada kelas minoritas dengan memanfaatkan informasi tetangga terdekat, sehingga memperkaya representasi kelas minoritas dan mengurangi bias model terhadap kelas mayoritas. Selanjutnya, ENN diterapkan untuk menghapus sampel yang bersifat ambigu atau berpotensi sebagai noise, baik dari kelas mayoritas maupun minoritas, berdasarkan ketidaksesuaian label dengan tetangga terdekatnya. Kombinasi kedua pendekatan ini tidak hanya menyeimbangkan distribusi kelas, tetapi juga meningkatkan kualitas data dengan membersihkan sampel yang dapat mengganggu proses pembelajaran. Dalam konteks dataset medis yang sering mengandung tumpang tindih antarkelas dan noise, seperti data klinis penyakit hati, SMOTE-ENN terbukti membantu meningkatkan kemampuan generalisasi model serta kinerja klasifikasi, khususnya pada metrik yang sensitif terhadap ketidakseimbangan data seperti *recall* dan *F1-score*. Namun demikian, karena ENN dapat mengurangi jumlah sampel secara signifikan, penerapan SMOTE-ENN perlu dilakukan secara hati-hati agar tidak menghilangkan informasi penting dari data asli. Setelah seluruh tahap preprocessing selesai, data telah siap digunakan untuk tahap pelatihan dan evaluasi model *machine learning*.



Gambar 2 Distribusi kelas

3.4 Model Machine Learning

3.4.1 Multilayer Perceptron (MLP)

MLP merupakan model *artificial neural network* berbasis *feed-forward* yang terdiri dari lapisan input, satu atau lebih lapisan tersembunyi (*hidden layers*), serta lapisan *output*. Setiap *neuron* dalam jaringan saling terhubung melalui bobot yang diperbarui menggunakan algoritma *backpropagation*. MLP memiliki kemampuan untuk mempelajari hubungan nonlinier yang kompleks antar fitur, sehingga banyak digunakan dalam berbagai permasalahan klasifikasi dan prediksi berbasis data numerik [20], [21]. Model MLP dipilih dalam penelitian ini karena karakteristik data klinis penyakit liver yang melibatkan hubungan nonlinier antara parameter laboratorium, faktor demografis, dan kondisi kesehatan pasien. Pendekatan berbasis jaringan saraf seperti MLP dinilai efektif dalam menangkap interaksi antar fitur yang sulit dimodelkan oleh metode linier konvensional [22].

 ISSN:2302-8149
e-ISSN:2540-9719

3.4.2 AdaBoost

AdaBoost (Adaptive Boosting) merupakan metode *ensemble learning* yang mengombinasikan sejumlah *weak learner* untuk membentuk sebuah *strong classifier*. Algoritma ini bekerja secara iteratif dengan memberikan bobot yang lebih besar pada sampel yang salah diklasifikasikan pada iterasi sebelumnya, sehingga model secara bertahap berfokus pada data yang lebih sulit diprediksi. Dengan mekanisme ini, AdaBoost mampu meningkatkan kinerja klasifikasi secara signifikan tanpa meningkatkan kompleksitas masing-masing model dasar [23], [24]. Dalam penelitian ini, algoritma AdaBoost diimplementasikan dengan decision tree berkedalaman satu (*decision stump*) sebagai *base estimator* atau *weak learner* [25]. Pemilihan konfigurasi ini sejalan dengan prinsip dasar AdaBoost yang menggabungkan sejumlah *weak learners* sederhana untuk membentuk *classifier* yang lebih kuat. Model AdaBoost dipilih karena kemampuannya dalam meningkatkan akurasi prediksi pada dataset penyakit *liver* yang memiliki pola nonlinier, kompleks, serta potensi ketidakseimbangan kelas, dan telah banyak diterapkan secara luas pada berbagai studi diagnosis penyakit berbasis data medis [26].

3.4.3 CatBoost

CatBoost (Categorical Boosting) merupakan algoritma *gradient boosting* berbasis *decision tree* yang dikembangkan untuk mengatasi keterbatasan metode *boosting* konvensional, khususnya terkait *overfitting* dan bias prediksi akibat urutan data (*prediction shift*). CatBoost membangun model secara bertahap dengan menambahkan pohon keputusan baru pada setiap iterasi untuk meminimalkan fungsi kerugian, sehingga mampu mempelajari hubungan nonlinier yang kompleks antar fitur secara efektif [27], [28]. Model CatBoost dipilih dalam penelitian ini karena kemampuannya menghasilkan performa klasifikasi yang stabil dan akurat pada data klinis, termasuk data penyakit *liver* yang memiliki pola kompleks dan potensi ketidakseimbangan kelas. Penggunaan struktur *symmetric tree* dan pendekatan *ordered boosting* memungkinkan CatBoost mengurangi risiko *overfitting* serta meningkatkan kemampuan generalisasi model, terutama pada dataset berukuran kecil hingga menengah yang umum dijumpai pada studi medis [29].

3.4.4 Pelatihan dan Arsitektur Model

Seluruh model *machine learning* dalam penelitian ini dilatih menggunakan data yang telah melalui tahap *preprocessing*. Untuk memperoleh konfigurasi hiperparameter yang optimal, dilakukan optimasi hiperparameter menggunakan metode GridSearchCV. *Grid search* merupakan pendekatan *exhaustive search* yang menguji seluruh kombinasi hiperparameter pada parameter grid yang telah ditentukan secara sistematis, kemudian memilih kombinasi dengan kinerja terbaik berdasarkan metrik evaluasi yang digunakan [30]. Proses ini dipadukan dengan cross-validation, yaitu teknik validasi internal yang membagi data pelatihan ke dalam beberapa fold sehingga model dilatih dan diuji secara bergantian pada setiap fold, menghasilkan estimasi performa yang lebih stabil dan lebih representatif terhadap kemampuan generalisasi dibandingkan satu kali pembagian train-test [31]. Dengan demikian, konfigurasi hiperparameter yang dipilih tidak bergantung pada satu pembagian data tertentu dan dapat mengurangi risiko bias evaluasi pada data klinis.

Proses pencarian hiperparameter mencakup variasi parameter-parameter utama yang berpengaruh terhadap kapasitas model, laju pembelajaran, serta tingkat kompleksitas struktur pembelajaran. Seluruh konfigurasi dievaluasi secara konsisten menggunakan skema validasi yang sama, sehingga perbandingan performa antar model dapat dilakukan secara adil dan objektif. Ringkasan parameter model, parameter GridSearchCV, serta pengaturan pelatihan untuk masing-masing algoritma disajikan secara lengkap pada Tabel 1. Seluruh proses optimasi hiperparameter pada penelitian ini menggunakan metrik ROC-AUC sebagai kriteria pemilihan model terbaik. Pemilihan metrik ROC-AUC didasarkan pada kemampuannya dalam mengevaluasi performa diskriminasi model secara menyeluruh pada berbagai ambang keputusan, sehingga lebih sesuai untuk dataset medis yang berpotensi mengalami ketidakseimbangan kelas [32]. Penggunaan metrik ini memungkinkan evaluasi yang lebih robust dibandingkan metrik berbasis akurasi semata, khususnya dalam konteks prediksi penyakit liver.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 Konfigurasi model

Model	Parameter Model	Parameter GridSearchCV	Pengaturan Pelatihan
MLP	- max_iter=300, - early_stopping=True, - random_state=42	- hidden_layer_sizes ∈ {(64), (128), (64,32), (128,64)}, - activation ∈ {ReLU, Tanh}, - solver ∈ {Adam, SGD}, - alpha ∈ {1e−4, 1e−3, 1e−2}, - learning_rate_init ∈ {0.001, 0.01}	10-fold Stratified Cross-Validation, scoring = AUROC
AdaBoost	- Base estimator: Decision Tree, - random_state=42	- n_estimators ∈ {50, 100, 200, 300}, - learning_rate ∈ {0.01, 0.1, 0.5, 1.0}, - max_depth (base tree) ∈ {1, 2, 3}	10-fold Stratified Cross-Validation, scoring = AUROC
CatBoost	- loss_function=Logloss, - eval_metric=AUC, - verbose=0, - random_state=42	- depth ∈ {4, 6, 8}, - learning_rate ∈ {0.01, 0.05, 0.1}, - iterations ∈ {200, 400, 800}, - l2_leaf_reg ∈ {1, 3, 5}	5-fold Stratified Cross-Validation, scoring = AUROC

3.5 Metrik Evaluasi

Untuk mengevaluasi kinerja model klasifikasi secara komprehensif, penelitian ini menggunakan beberapa metrik evaluasi yang umum digunakan dalam bidang *machine learning*, yaitu accuracy, precision, recall, F1-score, dan Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data, sehingga memberikan gambaran umum performa model. Precision mengukur tingkat ketepatan model dalam memprediksi kelas positif, yaitu proporsi prediksi positif yang benar-benar positif. Recall mengukur kemampuan model dalam mengidentifikasi seluruh sampel positif yang sebenarnya. F1-score merupakan rata-rata harmonik antara precision dan recall yang digunakan untuk menyeimbangkan kedua metrik tersebut, terutama pada kondisi ketidakseimbangan kelas. Selain itu, ROC-AUC digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai ambang (*threshold*). Nilai AUC yang mendekati 1 menunjukkan kemampuan diskriminasi model yang semakin baik, sedangkan nilai mendekati 0,5 menunjukkan performa yang setara dengan tebakan acak [33], [34], [35], [36]. Metrik evaluasi yang digunakan didefinisikan pada Persamaan 2-6.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{AUC} = \int_{\{0\}}^{\{1\}} TPR(FPR)\, d(FPR) \tag{6}$$

 ISSN:2302-8149
e-ISSN:2540-9719

Di mana *TP* merupakan *True Positive* (kasus positif yang diprediksi benar), *TN* merupakan *True Negative* (kasus negatif yang diprediksi benar), *FP* merupakan *False Positive* (kasus negatif yang diprediksi positif), dan *FN* merupakan *False Negative* (kasus positif yang diprediksi negatif). *True Positive Rate* (TPR) merepresentasikan proporsi data positif yang berhasil diklasifikasikan dengan benar oleh model, sehingga mencerminkan tingkat sensitivitas model. Sementara itu, *False Positive Rate* (FPR) menunjukkan proporsi data negatif yang secara keliru diklasifikasikan sebagai positif. Nilai ROC–AUC dihitung sebagai luas area di bawah kurva ROC yang menggambarkan hubungan antara TPR dan FPR pada berbagai nilai ambang keputusan.

4 Hasil dan Pembahasan

4.1 Hasil

Berdasarkan hasil evaluasi pada data uji yang dirangkum dalam Tabel 2, ketiga model menunjukkan performa yang kompetitif dengan karakteristik yang berbeda dalam menangani ketidakseimbangan kelas. Model MLP memberikan kinerja terbaik secara keseluruhan, dengan akurasi tertinggi sebesar 75% dan nilai ROC–AUC 0.848, yang mengindikasikan kemampuan diskriminasi kelas yang paling baik di antara seluruh model yang diuji. Pada kelas non-liver, MLP mencapai nilai recall yang sangat tinggi (0.94), menunjukkan kemampuannya dalam mengidentifikasi sampel negatif secara efektif, meskipun nilai precision relatif rendah (0.53). Sementara itu, pada kelas liver, MLP menghasilkan precision yang sangat tinggi (0.96) dengan recall 0.67, sehingga memperoleh F1-score 0.79, yang mencerminkan keseimbangan yang baik antara ketepatan prediksi dan sensitivitas terhadap pasien penyakit hati. Model AdaBoost menunjukkan performa yang lebih rendah dibandingkan MLP, dengan akurasi 71% dan nilai ROC–AUC 0.824. Pada kelas liver, AdaBoost mencapai precision 0.86 dan recall 0.70, yang menghasilkan F1-score 0.78, menandakan kemampuan moderat dalam mendeteksi pasien penyakit hati. Namun, performanya pada kelas *non-liver* relatif lebih rendah dibandingkan MLP, terutama pada aspek precision (0.50), yang berkontribusi terhadap akurasi keseluruhan yang lebih rendah.

Sementara itu, CatBoost menunjukkan kinerja yang stabil dengan akurasi 74% dan nilai ROC–AUC 0.826, sedikit lebih baik dibandingkan AdaBoost namun masih berada di bawah MLP. Pada kelas *non-liver*, CatBoost memperoleh recall tinggi (0.85) dengan precision 0.53, sedangkan pada kelas liver model ini menghasilkan precision tinggi (0.92) dengan recall 0.69, sehingga menghasilkan F1-score 0.79, sebanding dengan MLP. Hasil ini menunjukkan bahwa CatBoost mampu memberikan prediksi yang konsisten dan cukup efektif, khususnya dalam menjaga ketepatan prediksi pada kelas positif.

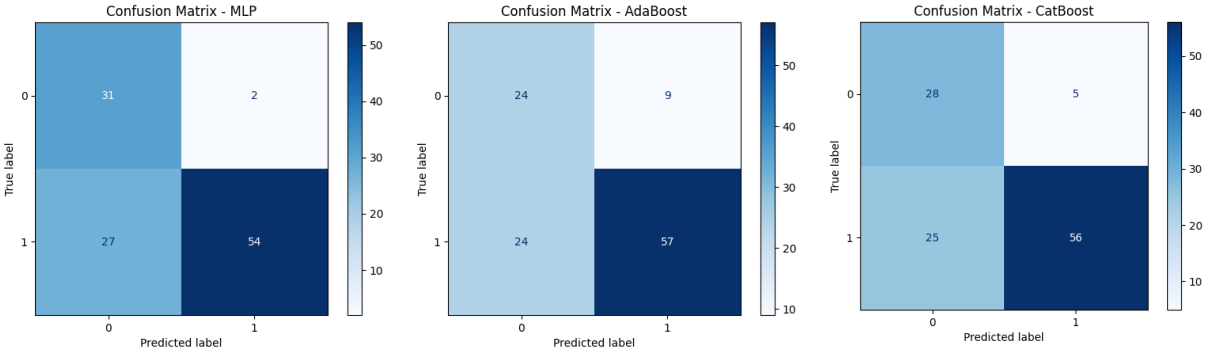
Tabel 2 Performa model

Model	Class	Precision	Recall	F1-Score	Accuracy	ROC-AUC
MLP	Non-liver	0.53	0.94	0.68	75%	0.848
	Liver	0.96	0.67	0.79		
AdaBoost	Non-liver	0.50	0.73	0.59	71%	0.824
	Liver	0.86	0.70	0.78		
CatBoost	Non-liver	0.53	0.85	0.65	74%	0.826
	Liver	0.92	0.69	0.79		

Gambar 3 menunjukkan *confusion matrix* dari masing-masing model. MLP mengklasifikasikan dengan benar 31 pasien *non-liver* dan 54 pasien *liver*, sedangkan 2 pasien *non-liver* dan 27 pasien *liver* salah diklasifikasikan. AdaBoost mengklasifikasikan dengan benar 24 pasien *non-liver* dan 57 pasien *liver*, dengan 9 dan 24 pasien dari masing-masing kelas salah diklasifikasikan. Sementara itu, CatBoost mengklasifikasikan dengan benar 28 pasien *non-liver* dan 56 pasien *liver*, sedangkan 5 pasien *non-liver* dan 25 pasien *liver* salah diklasifikasikan. Hasil ini menunjukkan bahwa ketiga model memiliki kemampuan identifikasi yang lebih baik pada kelas *liver* dibandingkan kelas *non-liver*. Kesalahan klasifikasi paling banyak berupa pasien *liver* yang diprediksi sebagai *non-liver*,

 ISSN:2302-8149
e-ISSN:2540-9719

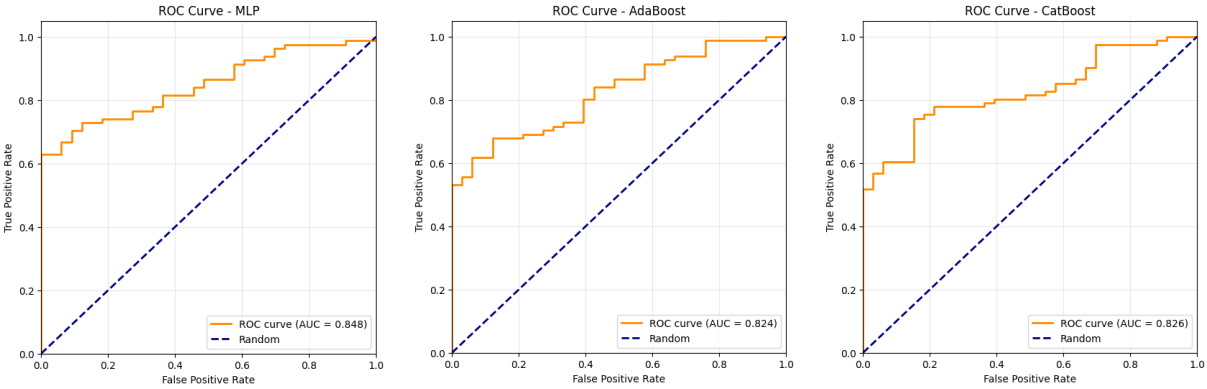
sehingga kemampuan model dalam mendeteksi pasien tanpa penyakit hati masih relatif lebih baik dibandingkan kemampuannya mengidentifikasi seluruh pasien yang benar-benar menderita penyakit hati.



(a) (b) (c)

Gambar 3 Confusion matrix: (a) MLP; (b) AdaBoost; (c) CatBoost

Gambar 4 menampilkan kurva ROC untuk model MLP, AdaBoost, dan CatBoost, yang menggambarkan kemampuan masing-masing model dalam membedakan pasien dengan dan tanpa penyakit *liver* pada berbagai ambang keputusan. Seluruh kurva ROC berada di atas garis diagonal acak, menunjukkan bahwa ketiga model memiliki kemampuan diskriminasi yang baik meskipun menghadapi potensi ketidakseimbangan kelas. Di antara ketiganya, model MLP menunjukkan performa paling unggul dengan kurva yang secara konsisten berada lebih tinggi pada sebagian besar rentang *false positive rate*, menandakan kemampuan yang lebih baik dalam mencapai sensitivitas tinggi tanpa peningkatan kesalahan klasifikasi yang signifikan. Model CatBoost dan AdaBoost juga memperlihatkan kurva ROC yang stabil dan kompetitif, meskipun berada sedikit di bawah MLP, yang mengindikasikan bahwa pendekatan *boosting* tetap efektif dalam menangkap pola klinis pada dataset ILPD.



(a) (b) (c)

Gambar 4 Kurva ROC: (a) MLP; (b) AdaBoost; (c) CatBoost

4.2 Perbandingan dengan Penelitian Terdahulu

Perbandingan kinerja model yang diusulkan dengan penelitian terdahulu pada dataset ILPD disajikan pada Tabel 3. Berbeda dengan sebagian besar penelitian sebelumnya yang hanya melaporkan metrik akurasi, penelitian ini menyajikan evaluasi yang lebih komprehensif dengan menambahkan metrik ROC–AUC untuk menilai kemampuan diskriminasi model secara lebih menyeluruh, khususnya dalam konteks data medis yang memiliki potensi ketidakseimbangan kelas.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 3 Perbandingan dengan penelitian sebelumnya

Studi	Dataset	Metode	Akurasi Terbaik	ROC-AUC
[10]	ILPD	Logistic Regression, Decision Tree, Random Forest, KNN, Gradient Boosting, XGBoost, LightGBM	63% (Random Forest dan LightGBM)	–
[11]	ILPD	Voting Ensemble (XGBoost, Bagging Meta-Estimator, MLP)	73.56%	–
[12]	ILPD	Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, k-NN, Logistic Regression	71.80% (Logistic Regression)	–
Penelitian ini	**ILPD**	**MLP, AdaBoost, CatBoost + GridSearchCV**	**75% (MLP)**	**0.848 (MLP)**

Secara umum, hasil penelitian ini menunjukkan tingkat akurasi yang lebih tinggi dibandingkan dengan studi-studi sebelumnya yang menggunakan dataset ILPD. Peningkatan kinerja tersebut dapat dikaitkan dengan penerapan strategi optimasi hiperparameter yang sistematis menggunakan GridSearchCV, sehingga setiap model dilatih pada konfigurasi yang paling sesuai dengan karakteristik data. Selain itu, penggunaan skema cross-validation dalam proses pelatihan dan pemilihan model membantu mengurangi bias akibat pembagian data tunggal serta meningkatkan kemampuan generalisasi model. Pemilihan metrik ROC–AUC sebagai kriteria evaluasi utama juga berperan dalam mendorong pemilihan model dengan kemampuan diskriminasi kelas yang lebih baik, terutama pada kondisi ketidakseimbangan kelas yang umum dijumpai pada data klinis. Kombinasi dari faktor-faktor tersebut berkontribusi terhadap peningkatan akurasi dan stabilitas performa model dibandingkan dengan pendekatan yang digunakan pada penelitian-penelitian sebelumnya.

4.3 Pembahasan

Analisis hasil menunjukkan bahwa kombinasi preprocessing yang tepat, seleksi fitur, dan strategi penyeimbangan kelas memengaruhi kinerja model. Robust Scaler membantu mengurangi pengaruh *outlier*, sementara RFE dengan Random Forest efektif memilih fitur paling relevan, sehingga model dapat mempelajari pola signifikan tanpa terganggu oleh variabel yang kurang informatif. Penerapan SMOTE-ENN pada data pelatihan menyeimbangkan distribusi kelas dan mengurangi sampel ambigu, yang mendukung peningkatan recall pada kelas minoritas, hal tersebut penting dalam konteks medis agar pasien dengan penyakit hati lebih mudah terdeteksi.

Dari sisi algoritma, MLP menonjol dalam menangkap hubungan nonlinier antar fitur, sedangkan CatBoost menunjukkan prediksi yang stabil pada kedua kelas. AdaBoost memberikan performa yang kompetitif, meskipun sedikit kurang stabil pada kelas *non-liver*. Selain itu, fitur-fitur yang terpilih melalui RFE memberikan *insight* mengenai variabel yang paling berpengaruh terhadap prediksi penyakit hati, yang dapat berguna untuk interpretasi klinis dan analisis lebih lanjut. Temuan ini menegaskan bahwa integrasi preprocessing, seleksi fitur, dan penyeimbangan kelas merupakan strategi efektif untuk meningkatkan kemampuan generalisasi model pada dataset medis yang tidak seimbang.

5 Kesimpulan

Berdasarkan penelitian ini, kombinasi preprocessing yang tepat, seleksi fitur menggunakan RFE, dan penyeimbangan kelas melalui SMOTE-ENN berhasil meningkatkan kinerja model dalam memprediksi pasien dengan penyakit hati pada dataset ILPD. 8 fitur terpilih, yaitu Age, TB, DB, Alkphos, Sgpt, Sgot, TP, dan ALB terbukti paling berpengaruh terhadap prediksi penyakit hati. Dari ketiga model yang diuji, MLP menunjukkan kemampuan diskriminasi terbaik secara keseluruhan,

CatBoost menawarkan prediksi yang lebih stabil, dan AdaBoost tetap memberikan performa kompetitif meskipun sedikit kurang stabil pada kelas non-liver. Hasil ini menunjukkan bahwa pendekatan yang diusulkan efektif dan dapat diandalkan untuk prediksi penyakit hati. Meskipun demikian, penelitian ini masih terbatas oleh ukuran dan ketidakseimbangan dataset. Pengembangan di masa mendatang dapat difokuskan pada pemanfaatan dataset klinis yang lebih besar dan beragam, penambahan fitur klinis relevan, serta eksplorasi model ensemble atau hybrid untuk meningkatkan ketahanan prediksi dan kinerja dalam sistem pendukung keputusan klinis.

Referensi

[1] C. Gan *et al.*, "*Liver Diseases: Epidemiology, Causes, Trends and Predictions*," *Signal Transduct. Target. Ther.*, Vol. 10, No. 1, p. 33, Feb. 2025, DOI: 10.1038/s41392-024-02072-z.

[2] R.-H. Lin, "*An Intelligent Model for Liver Disease Diagnosis*," *Artif. Intell. Med.*, Vol. 47, No. 1, pp. 53–62, Sep. 2009, DOI: 10.1016/j.artmed.2009.05.005.

[3] S. G. Sepanlou *et al.*, "*The Global, Regional, and National Burden of Cirrhosis by Cause in 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017*," *Lancet Gastroenterol. Hepatol.*, Vol. 5, No. 3, pp. 245–266, Mar. 2020, DOI: 10.1016/S2468-1253(19)30349-8.

[4] E. B. Tapper and N. D. Parikh, "*Mortality Due to Cirrhosis and Liver Cancer in the United States, 1999-2016: Observational Study*," *BMJ*, p. k2817, Jul. 2018, DOI: 10.1136/bmj.k2817.

[5] K. R. Islam *et al.*, "*Machine Learning-based Early Prediction of Sepsis using Electronic Health Records: A Systematic Review*," *J. Clin. Med.*, Vol. 12, No. 17, p. 5658, Aug. 2023, DOI: 10.3390/jcm12175658.

[6] H. Sadr *et al.*, "*Unveiling the Potential of Artificial Intelligence in Revolutionizing Disease Diagnosis and Prediction: A Comprehensive Review of Machine Learning and Deep Learning Approaches*," *Eur. J. Med. Res.*, Vol. 30, No. 1, p. 418, May 2025, DOI: 10.1186/s40001-025-02680-7.

[7] H. Abualruz *et al.*, "*The Role of Artificial Intelligence in Enhancing Triage Decisions in Healthcare Settings: A Systematic revIew*," *Applied Nursing Research*, Vol. 86, p. 152024, Dec. 2025, DOI: 10.1016/j.apnr.2025.152024.

[8] S. Shen *et al.*, "*Mapping the Landscape of Machine Learning in Chronic Disease Management: A Comprehensive Bibliometric Study*," *Digit. Health*, Vol. 11, May 2025, DOI: 10.1177/20552076251361614.

[9] A. P. Susanto, D. Lyell, B. Widyantoro, S. Berkovsky, and F. Magrabi, "*Effects of Machine Learning-based Clinical Decision Support Systems on Decision-Making, Care Delivery, and Patient Outcomes: A Scoping Review*," *Journal of the American Medical Informatics Association*, Vol. 30, No. 12, pp. 2050–2063, Nov. 2023, DOI: 10.1093/jamia/ocad180.

[10] K. Gupta, N. Jiwani, N. Afreen, and D. D, "*Liver Disease Prediction using Machine Learning Classification Techniques*," in *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, Apr. 2022, pp. 221–226. DOI: 10.1109/CSNT54456.2022.9787574.

[11] I. Altaf, M. A. Butt, and M. Zaman, "*Hard Voting Meta Classifier for Disease Diagnosis using Mean Decrease in Impurity for Tree Models*," *Review of Computer Engineering Research*, Vol. 9, No. 2, pp. 71–82, Jun. 2022, DOI: 10.18488/76.v9i2.3037.

[12] I. R. Hikmah and R. N. Yasa, "Perbandingan Hasil Prediksi Diagnosis pada *Indian Liver Patient Dataset (ILPD)* dengan Teknik *Supervised Learning* menggunakan *Software Orange*," *Jurnal Telematika*, Vol. 16, No. 2, pp. 69–76, Jan. 2022, DOI: 10.61769/telematika.v16i2.402.

[13] A. Agung Gde Wahyu Sukma Erlangga, P. Putu Ode Juliantara, and Adnan, "Perbandingan Performa *Smote* dan *Smote-Enn* pada Data Tidak Seimbang," *J. RAMATEKNO*, Vol. 5, No. 2, pp. 150–158, 2025.

[14] 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. (Lausanne).*, Vol. 12, May 2025, DOI: 10.3389/fmed.2025.1502749.

 ISSN:2302-8149
e-ISSN:2540-9719

[15] B. Venkata Ramana, M. Surendra Prasad Babu, N. B. Venkateswarlu, and A. Professor, "*A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis*," 2012. [Online]. Available: www.IJCSI.org

[16] J. M. Kernbach and V. E. Staartjes, "*Foundations of Machine Learning-based Clinical Prediction Modeling: Part II—Generalization and Overfitting*," 2022, pp. 15–21. DOI: 10.1007/978-3-030-85292-4_3.

[17] E. Yalçın, T. K. Koç, S. Aslan, S. C. Demir, S. Aykut, and M. Sucu, "*Hyperparameter Optimization based Machine Learning Approach for Early Diagnosis of Fetal Genetic Disorders*," *Discover Computing*, Vol. 28, No. 1, p. 274, Nov. 2025, DOI: 10.1007/s10791-025-09815-8.

[18] X. Chen and J. C. Jeong, "*Enhanced Recursive Feature Elimination*," in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, IEEE, Dec. 2007, pp. 429–435. DOI: 10.1109/ICMLA.2007.35.

[19] X. Zhou, H. Wen, Y. Zhang, J. Xu, and W. Zhang, "*Landslide Susceptibility Mapping using Hybrid Random Forest with GeoDetector and RFE for Factor Optimization*," *Geoscience Frontiers*, Vol. 12, No. 5, p. 101211, Sep. 2021, DOI: 10.1016/j.gsf.2021.101211.

[20] F. Dashti, A. Ghaffari, A. Seyfollahi, and B. Arasteh, "*A Self-Predictive Diagnosis System of Liver Failure based on Multilayer Neural Networks*," *Multimed. Tools Appl.*, Vol. 83, No. 36, pp. 83769–83788, Mar. 2024, DOI: 10.1007/s11042-024-18945-y.

[21] A. Singh, S. Kushwaha, M. Alarfaj, and M. Singh, "*Comprehensive Overview of Backpropagation Algorithm for Digital Image Denoising*," *Electronics (Basel).*, Vol. 11, No. 10, p. 1590, May 2022, DOI: 10.3390/electronics11101590.

[22] 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.

[23] Y. Wang and L. Feng, "*An Adaptive Boosting Algorithm based on Weighted Feature Selection and Category Classification Confidence*," *Applied Intelligence*, Vol. 51, No. 10, pp. 6837–6858, Oct. 2021, DOI: 10.1007/s10489-020-02184-3.

[24] Q. Sui and S. K. Ghosh, "*Active Learning for Stacking and AdaBoost-Related Models*," *Stats (Basel).*, Vol. 7, No. 1, pp. 110–137, Jan. 2024, DOI: 10.3390/stats7010008.

[25] M. Bereta, "*Regularization of Boosted Decision Stumps using Tabu Search*," *Appl. Soft Comput.*, Vol. 79, pp. 424–438, Jun. 2019, DOI: 10.1016/j.asoc.2019.04.003.

[26] J. Hatwell, M. M. Gaber, and R. M. Atif Azad, "*Ada-WHIPS: Explaining AdaBoost Classification with Applications in the Health Sciences*," *BMC Med. Inform. Decis. Mak.*, Vol. 20, No. 1, p. 250, Dec. 2020, DOI: 10.1186/s12911-020-01201-2.

[27] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "*CatBoost: Unbiased Boosting with Categorical Features*," Jan. 2019.

[28] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, "*CatBoost Model and Artificial Intelligence Techniques for Corporate Failure Prediction*," *Technol. Forecast. Soc. Change*, Vol. 166, p. 120658, May 2021, DOI: 10.1016/j.techfore.2021.120658.

[29] 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.

[30] D. El-Shahat, A. Tolba, M. Abouhawwash, and M. Abdel-Basset, "*Machine Learning and Deep Learning Models based Grid Search Cross Validation for Short-Term Solar Irradiance Forecasting*," *J. Big Data*, Vol. 11, No. 1, p. 134, Sep. 2024, DOI: 10.1186/s40537-024-00991-w.

[31] S. Bates, T. Hastie, and R. Tibshirani, "*Cross-Validation: What Does it Estimate and How Well Does It do It?*," Jul. 2022, DOI: 10.1080/01621459.2023.2197686.

[32] 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.

[33] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "*Accuracy, Precision, Recall, f1-Score, or MCC? Empirical Evidence from Advanced Statistics, ML, and XAI for Evaluating Business*

 ISSN:2302-8149
e-ISSN:2540-9719

Predictive Models," *J. Big Data*, Vol. 12, No. 1, p. 268, Dec. 2025, DOI: 10.1186/s40537-025-01313-4.

[34] O. Rainio, J. Teuho, and R. Klén, "*Evaluation Metrics and Statistical Tests for Machine Learning*," *SCI. Rep.*, Vol. 14, No. 1, p. 6086, Mar. 2024, Doi: 10.1038/s41598-024-56706-x.

[35] D. M. W. Powers, "*Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*," Oct. 2020.

[36] A. P. Bradley, "*The use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms*," *Pattern Recognit.*, Vol. 30, No. 7, pp. 1145–1159, Jul. 1997, DOI: 10.1016/S0031-3203(96)00142-2.