Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

Implementation of Extra Trees Classifier and Chi-Square Feature Selection for Early Detection of Liver Disease

Muhammad Akmal Al Ghifari[1], Irwan Budiman[*2], Triando Hamonangan Saragih[3], Muhammad Itqan Mazdadi[4], Rudy Herteno[5], Hasri Akbar Awal Rozaq[6]

[1,2,3,4,5]Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Indonesia
[6]Graduate School of Informatics, Department of Computer Science, Gazi University, Türkiye

Email: 2irwan.budiman@ulm.ac.id

Received : Dec 27, 2024; Revised : Feb 10, 2025; Accepted : Feb 19, 2025; Published : Oct 16, 2025

Abstract

The imbalanced distribution of medical data poses challenges in accurately detecting liver disease, which is crucial as symptoms often remain unnoticed until advanced stages. This study examines the application of the Extra Trees Classifier algorithm and chi-square feature selection for early detection of liver disease. Compared to traditional methods like Random Forest and SVM, the Extra Trees Classifier offers enhanced computational efficiency and better handling of imbalanced datasets, while chi-square feature selection helps identify the most relevant medical indicators. The data consists of five medical variables likely to be laboratory test results from patient samples, with labels indicating classes A and B. The data is randomly divided with a ratio of 80% for each class. To address data imbalance, SMOTE technique was applied before the data was randomly split into a ratio of 80% for training and 20% for testing to ensure effective learning and testing of the model's performance. The results showed that with the help of chi-square feature selection, the Extra Trees Classifier algorithm could provide fairly accurate predictions in liver disease classification, with an accuracy of 82.6%, sensitivity of 85.5%, precision of 78.3%, and F1-Score of 81.7%. These results demonstrate significant improvement over existing methods, and the proposed approach can aid healthcare practitioners in making timely diagnostic decisions, potentially reducing mortality rates through early intervention in liver disease cases.

Keywords : *Chi-Square, Extra Trees Classifier, Feature Selection, Liver.*

1. PENDAHULUAN

Liver atau hati adalah organ yang vital bagi manusia. Organ ini terletak di dalam rongga perut sebelah kanan, tepatnya di bawah diafragma [1]. Terdapat beberapa fungsi kerja liver antara lain sebagai penawar dan penetralisir racun, mengatur sirkulasi hormon, mengatur komposisi darah yang mengandung lemak, gula, protein, dan zat lain [2]. Liver juga berfungsi membuat empedu, zat yang membantu pencernaan lemak [3]. Penyakit liver merupakan suatu gangguan pada setiap fungsi hati [4]. Penyakit liver sering disebut sebagai pembunuh diam, karena kemungkinan tidak timbulnya gejala [5]. Penyakit liver merupakan penyakit peradangan pada organ hati, secara umum faktor penyebab terjadinya penyakit liver dapat disebabkan oleh pola hidup yang tidak sehat namun faktor lainnya adalah kondisi adanya kelainan hati yang merupakan bawaan sejak lahir atau pada saat kelahiran, adanya gangguan dan kelainan pada proses metabolisme, terinfeksi virus atau bakteri, kekurangan gizi atau nutrisi, ketergantungan alkohol dan zat adiktif lainnya maupun kecanduan dan kebiasaan merokok juga dapat menjadi penyebab dari penyakit liver [6].

Permasalahan yang dihadapi masyarakat saat ini salah satunya adalah keterlambatan penanganan secara medis kepada penderita liver karena sebagian besar pasien memeriksakan kondisinya setelah

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

penyakit terdeteksi sudah pada stadium lanjut [7]. Keterlambatan diagnosis penyakit liver dapat menyebabkan berbagai komplikasi serius seperti sirosis hati, kanker hati, dan bahkan kematian. Penelitian menunjukkan bahwa tingkat kematian akibat penyakit liver meningkat hingga 50% pada kasus yang terdiagnosis terlambat [8]. Untuk menanggulangi masalah terjadi semakin parahnya kondisi kesehatan penderita maka diperlukan pemeriksaan rutin dan pencegahan risiko adanya serangan penyakit kronis tersebut. Namun, pemeriksaan rutin dan pencegahan risiko terhadap penyakit liver tersebut tidak dilakukan oleh sebagian masyarakat karena beberapa alasan di antaranya rutinitas yang padat, mahalnya biaya pemeriksaan serta takut akan adanya diagnosa penyakit kronis [9]. Pemeriksaan rutin dan pencegahan risiko adanya gejala penyakit liver sejak dini ini sangat diperlukan, agar penderita dapat melakukan pengobatan dengan tepat. Dengan diagnosa adanya penyakit liver lebih awal dapat meningkatkan harapan hidup pasien. Meningkatkan harapan hidup pasien yang menderita penyakit liver dapat meningkat jika diagnosis awal dilakukan. Untuk mendiagnosis penyakit liver, tenaga ahli medis perlu melakukan banyak tes dan pemeriksaan untuk memastikan diagnosis tersebut, tetapi mereka tidak dapat memastikan kebenaran diagnosis tersebut. Salah satu tes dan pemeriksaan diagnosis penyakit liver adalah tes fungsi hati [2]. Tes fungsi hati sangat membantu dalam diagnosis penyakit liver. Parameter pengukuran dalam tes fungsi hati yang dilakukan di antaranya adalah albumin, alkali fosfatase, protein total, aspartat aminotransferase, alanine aminotransferase, bilirubin langsung, bilirubin total, gamma-glutamil transferase, waktu protrombin, jumlah trombosit, trigliserida, dan lainnya.

Berdasarkan dengan adanya kebutuhan untuk menemukan metode yang lebih akurat untuk mendeteksi dan mendiagnosis penyakit liver, peneliti mengkaji sebuah penerapan algoritma machine learning, yaitu algoritma Extra Trees Classifier serta penggunaan seleksi fitur chi-square untuk implementasi pendeteksi penyakit liver sejak dini [10], [11], [12], dengan tujuan untuk mendapatkan informasi tentang diagnosis lebih akurat serta untuk meningkatkan analisis tenaga ahli medis dalam mendiagnosis penyakit liver lebih awal [13]. Extra Trees Classifier atau bisa juga disebut Extremely Randomized Trees adalah jenis teknik machine learning ansambel yang menggabungkan hasil dari beberapa pohon keputusan yang tidak berkorelasi yang dikumpulkan pada pohon keputusan untuk mengeluarkan hasil klasifikasinya [14]. Sedangkan seleksi fitur chi-square adalah uji kebebasan dari dua variabel kualitatif untuk mengetahui apakah adanya keterkaitan antara dua variabel kategori [15]. Dengan kata lain, uji ini digunakan untuk mengetahui apakah nilai-nilai dari satu variabel kategori saling bergantung pada nilai-nilai variabel kategori yang lain.

Penerapan algoritma Extra Trees Classifier serta penggunaan seleksi fitur chi-square dalam bidang perawatan kesehatan sangat penting karena penerapan kedua metode tersebut dapat memprediksi pola di seluruh kumpulan data, yang memungkinkan penentuan risiko atau faktor diagnostik untuk penyakit dengan lebih cepat dan tepat [16]. Metode ini dapat memungkinkan deteksi dini dan mencegah banyak kasus penyakit liver berkembang hingga memerlukan biopsi atau pengobatan kompleks [17]. Jika penyakit liver terdiagnosis pada tahap awal, penyakit liver tersebut dapat ditangani. Bidang kesehatan banyak menggunakan teknik machine learning [18], terutama untuk diagnosis dan klasifikasi penyakit tertentu berdasarkan informasi karakteristik. Tenaga ahli medis akan lebih mudah membuat keputusan tentang pasien dengan menggunakan metode ini. Ketika data diperoleh menggunakan teknik pembuatan fitur machine learning, ruang fitur mentah masukan biasanya penuh dengan banyak informasi fitur yang tidak relevan dan sering kali menunjukkan dimensi tinggi.

Beberapa penelitian sebelumnya telah menunjukkan efektivitas machine learning dalam mendiagnosis penyakit liver. Panwar et al., (2022) telah melakukan penelitian dan menemukan hasil bahwa algoritma klasifikasi sering digunakan untuk meramalkan penyakit liver karena mereka dapat memprediksi apakah pasien memiliki penyakit atau tidak berdasarkan fitur atau karakteristik tertentu [19]. Penelitian lain yang dilakukan oleh Shaker Abdalrada et al., (2022), menemukan bahwa penggunaan kemampuan machine learning untuk membuat model prediksi yang berguna akan sangat

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

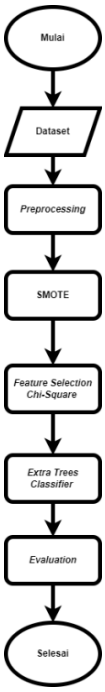
membantu dalam pengenalan penyakit dan pengambilan keputusan medis yang efektif secara real-time [20].

Meskipun berbagai metode machine learning telah digunakan untuk deteksi penyakit liver, seperti Random Forest, SVM, dan XGBoost, metode-metode tersebut memiliki beberapa keterbatasan. Random Forest cenderung mengalami overfitting pada dataset yang tidak seimbang, SVM membutuhkan waktu komputasi yang lama untuk dataset besar, dan XGBoost sensitif terhadap noise dan outlier [11], [21], [22]. Extra Trees Classifier mengatasi keterbatasan ini dengan beberapa keunggulan: (1) lebih tahan terhadap overfitting karena proses randomisasi yang lebih ekstrem, (2) waktu komputasi yang lebih cepat karena tidak perlu mencari split point optimal, dan (3) performa yang lebih baik pada dataset tidak seimbang [23], [24]. Selain itu, penggunaan seleksi fitur chi-square memberikan keuntungan dalam mengidentifikasi fitur-fitur medis yang paling relevan, mengurangi kompleksitas model, dan meningkatkan interpretabilitas hasil diagnosis [25].

Berdasarkan paparan permasalahan tersebut, penelitian ini akan menggunakan algoritma Extra Trees Classifier dan chi-square dalam pemilihan fitur untuk mengidentifikasi klasifikasi penyakit liver. Kombinasi kedua metode ini masih jarang dieksplorasi dalam konteks deteksi penyakit liver, padahal memiliki potensi untuk meningkatkan akurasi diagnosis sambil mempertahankan efisiensi komputasi. Penelitian ini akan menghasilkan akurasi yang tinggi dengan tujuan deteksi dini dan mengurangi kemungkinan kesalahan dalam pengenalan penyakit liver. Selain itu, penelitian ini akan dibandingkan dengan hasil penelitian sebelumnya untuk memberikan konteks yang lebih baik tentang efektivitas metode yang digunakan.

2. BAHAN DAN METODE

Dalam penelitian ini menggunakan salah satu teknik machine learning terkhususnya supervised learning yaitu metode klasifikasi dengan algoritma Extra Trees Classifier. Dalam penelitian ini beberapa tahapan dapat dilihat pada alur penelitian berikut.



Gambar 1. Alur Penelitian

Pada Gambar 1 di atas, data yang dikumpulkan dalam sebuah dataset akan dimasukkan ke tahap preprocessing, yang mencakup penanganan nilai yang hilang (missing value), normalisasi data, dan

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

pengkodean label. Setelah tahap preprocessing selesai, SMOTE digunakan untuk menyeimbangkan jumlah data kelas minoritas dengan data kelas mayoritas. Langkah berikutnya adalah pemilihan fitur menggunakan uji chi-square untuk menentukan fitur yang paling relevan. Selanjutnya, model dibangun menggunakan algoritma Extra Trees Classifier, dan tahap terakhir adalah evaluasi kinerja model menggunakan metrik yang relevan untuk memastikan akurasi dan efektivitas dalam mendeteksi penyakit liver.

2.1. Dataset

Dataset yang digunakan dalam penelitian ini adalah Indian Liver Patient Dataset, yang diperoleh dari UCI Machine Learning Repository dan dapat diakses melalui https://www.kaggle.com/datasets/uciml/indian-liver-patient-records. Dataset ini dipilih berdasarkan kelengkapan atribut yang relevan dengan tujuan penelitian. Proses seleksi dilakukan dengan mempertimbangkan kualitas data, termasuk validasi terhadap data yang hilang atau tidak konsisten. Data yang digunakan dalam analisis telah melalui tahap pembersihan dan preprocessing untuk memastikan keakuratan serta relevansi informasi yang diperoleh [26].

Dataset ini berisi rekam medis pasien yang terdiri dari 10 fitur dan 583 baris data, dengan dua label kelas: 416 baris untuk pasien dengan penyakit liver (kelas 1) dan 167 baris untuk pasien tanpa penyakit liver (kelas 2). Dataset dibagi menjadi dua bagian, yaitu data uji dan data latih, dengan rasio perbandingan 80% dan 20%. Pembagian rasio tersebut terbukti baik digunakan untuk proses analisis menggunakan machine learning [27]. Data Rincian masing-masing atribut data dapat dilihat pada tabel berikut.

Tabel 1. Deskripsi Fitur Dataset

Features	Data Type	Description
Age	Numerical	Attribute
Gender	Binary (0,1)	Attribute
Total Bilirubin	Numerical	Attribute
Direct Bilirubin	Numerical	Attribute
Alkaline_Phospotase	Numerical	Attribute
Alamine_Aminotransferase	Numerical	Attribute
Aspartate_Aminotransferase	Numerical	Attribute
Total_Protein	Numerical	Attribute
Albumin	Numerical	Attribute
Albumin_And_Globulin Ratio	Numerical	Attribute
Category	Numerical	Class

Pada Tabel 1, ditampilkan deskripsi rinci dari fitur-fitur yang terdapat dalam dataset penelitian. Tabel tersebut mencakup beberapa kolom. Age merupakan fitur numerik yang menunjukkan usia pasien, sedangkan Gender adalah fitur biner dengan nilai 0 atau 1 yang menunjukkan jenis kelamin pasien. Fitur Total Bilirubin mengukur total kadar bilirubin dalam darah dan merupakan fitur numerik, sama seperti Direct Bilirubin yang mengukur kadar bilirubin langsung dalam darah. Fitur Alkaline Phosphatase mengukur kadar enzim alkaline phosphatase dalam darah, dan Alamine Aminotransferase mengukur kadar enzim alamine aminotransferase dalam darah. Aspartate Aminotransferase adalah fitur numerik yang mengukur kadar enzim aspartate aminotransferase dalam darah, sementara Total Protein mengukur total kadar protein dalam darah. Albumin merupakan fitur numerik yang mengukur kadar albumin dalam darah, dan Albumin and Globulin Ratio mengukur rasio albumin dan globulin dalam darah. Terakhir,

Category adalah fitur numerik yang berfungsi sebagai kelas target atau variabel terikat, dengan dua kategori yang menunjukkan apakah pasien menderita penyakit liver atau tidak.

2.2. Pra-Pemrosesan Data

Pre-processing data adalah tahap penting yang harus dilakukan terlebih dahulu. Ini mengubah data mentah menjadi data berkualitas yang dapat diproses pada tahap berikutnya. Beberapa tahapan yang dilakukan dalam proses ini yaitu penanganan missing value, normalisasi data dan label encoding (Kurniawan dan Mustikasari, 2021).

Tahap pertama missing value adalah salah satu dari beberapa tantangan kualitas data yang sering terjadi dalam kumpulan data dunia nyata [29]. Masalah umum ini biasanya berdampak pada kinerja analitik data, menyebabkan bias tinggi dan akurasi rendah. Dalam penelitian ini, proses penanganan missing value adalah dengan mengganti nilai yang tidak tersedia dengan nilai rata-rata dari atribut yang relevan.

Tahap kedua normalisasi data diperlukan hanya ketika himpunan data memiliki rentang yang berbeda [30]. Tujuan normalisasi data adalah untuk mengubah nilai kolom numerik dalam himpunan data ke skala umum tanpa mengubah rentang nilai secara keseluruhan (Patimah, Haekal dan Sandya Prasvita, 2021). Proses normalisasi data adalah dengan menyamakan skala nilai antar variabel, yang juga meningkatkan akurasi karena dengan nilai yang sama maka model akan mengenali data dengan lebih efisien. Berikut rumus dasar penskalaan min-max seperti pada formula.

$$X_{norm} = \frac{x' - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Dimana x merupakan nilai atribut, min(x) dan max(x) adalah nilai absolut minimal dan maksimal dari x, x' adalah nilai lama dari setiap entri dalam data.

Tahap ketiga label encoding adalah metode yang digunakan dalam pemrosesan data untuk mengubah label atau kategori menjadi representasi numerik untuk menganalisis dan membuat model machine learning [31]. Tujuan melakukan label encoding adalah untuk membuat proses data mining mudah karena metode data mining umumnya lebih mampu membaca data bertipe angka.

2.3. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) adalah teknik yang digunakan dalam machine learning untuk mengatasi masalah ketidakseimbangan kelas dalam dataset [32]. Ketidakseimbangan kelas terjadi ketika satu kelas (minoritas) memiliki jauh lebih sedikit sampel dibandingkan kelas lainnya (mayoritas). Hal ini dapat menyebabkan model machine learning menjadi bias terhadap kelas mayoritas dan kurang efektif dalam memprediksi kelas minoritas. Ketidakseimbangan kelas dapat menyebabkan masalah signifikan dalam machine learning, seperti, model bias yaitu cenderung memprediksi kelas mayoritas lebih sering daripada kelas minoritas karena kelas mayoritas lebih terwakili dalam data pelatihan, kurangnya generalisasi dimana model tidak dapat mengenali pola yang cukup dalam kelas minoritas, yang dapat menyebabkan prediksi yang buruk ketika dihadapkan pada data baru yang mengandung kelas minoritas (Özdemir, Polat dan Alhudhaif, 2021). SMOTE bekerja dengan cara menghasilkan sampel sintetis dari kelas minoritas untuk meningkatkan jumlah sampelnya [34].

SMOTE meningkatkan jumlah baris data dengan menghasilkan data sintetis acak untuk kelas minoritas dari tetangga terdekatnya. Karena mereka dibuat berdasarkan karakteristik asli dataset, baris data baru lebih mirip dengan data asli [35]. Berikut adalah rumus dari perhitungan SMOTE menggunakan jarak geometri.

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

$$D(x,y) = \sqrt{(X_1 - Y_1)^2 + .. + (X_n - Y_n)^2} \quad (2)$$

Dalam persamaan ini, D(x,y) adalah jarak Euclidean antara dua titik, x dan y. Titik x merepresentasikan koordinat titik pertama dalam ruang n-dimensi, dengan n komponen yang dapat ditulis sebagai (X1, X2,…,Xn). Sedangkan titik y mewakili koordinat titik kedua dalam ruang n-dimensi, juga dengan n komponen yang ditulis sebagai (Y1,Y2,…,Yn). Komponen Xi adalah komponen ke-i dari titik x, di mana i berkisar dari 1 hingga n, dan Yi adalah komponen ke-i dari titik y, di mana i juga berkisar dari 1 hingga n. Setelah menghitung jarak instance dengan jarak geometri, formula (3) digunakan untuk membuat data replikasi dari instance terdekat.

$$Xsyn = X_i + (X_{knn} - X_i) \times \sigma \quad (3)$$

Dalam rumus ini, Xsyn merupakan titik data sintetis yang dihasilkan, Xi adalah titik data asli dari kelas minoritas yang dipilih, dan Xknn adalah titik data tetangga terdekat dari Xi dalam kelas minoritas. Variabel σ adalah nilai acak antara 0 dan 1. Proses pembentukan titik data sintetis dilakukan dengan memilih titik data asli X i dari kelas minoritas dan titik tetangga terdekatnya Xknn, kemudian mengambil selisih antara Xknn dan Xi, mengalikan selisih ini dengan nilai acak σ, dan menambahkan hasilnya ke Xi. Proses ini diulang untuk setiap titik data dalam kelas minoritas hingga dataset seimbang.

2.4. Seleksi Fitur

Seleksi fitur adalah proses dalam analisis data yang mengurangi dimensi data, meningkatkan efisiensi komputasi, menghilangkan atribut yang tidak memberikan kontribusi signifikan terhadap tugas analisis atau pemodelan, dan mencegah overfitting [32]. Data yang digunakan bersifat kategorikal, dimana pemilihan fitur dapat dilakukan dengan menggunakan metode chi-square.

Metode chi-square dalam pemilihan fitur merupakan salah satu teknik yang digunakan untuk mengukur hubungan antara atribut kategori dalam suatu dataset dan variabel target kategori. Ini membantu mengidentifikasi fitur yang memiliki hubungan kuat dengan variabel target dan dapat digunakan untuk memprediksi atau menjelaskan variabel target. Metode ini umumnya digunakan untuk masalah klasifikasi yang variabel sasarannya adalah variabel kategori atau kelas. Berikut rumus chi-square seperti yang ditunjukkan pada formula.

$$X^2 = [\frac{\Sigma(f_{(0)} - f_{(e)})^2}{f_{(0)}}] \quad (4)$$

Dimana X^2 adalah nilai chi-square, f_{(e)} adalah frekuensi yang diharapkan, dan f_{(0)} adalah frekuensi yang diperoleh/diamati.

2.5. *Extra Tree-Classifier*

Extra Trees Classifier adalah algoritma machine learning ansambel yang digunakan untuk klasifikasi. Ini termasuk dalam keluarga algoritma [32] Random Forest, yang membentuk sejumlah pohon keputusan dan menggunakan prediksi mereka untuk menghasilkan hasil akhir. Extra Trees Classifier berbeda dari Random Forest tradisional karena menggunakan ambang acak untuk setiap fitur daripada memilih pemisahan terbaik untuk membentuk setiap pohon keputusan.

Algoritma Extra Trees Classifier menggunakan prosedur pemisahan acak untuk atribut numerik. Prosedur ini diatur oleh parameter jumlah atribut yang dipilih secara acak untuk setiap node dan ukuran sampel minimum untuk pemisahan node. Metode ini digunakan untuk membuat model ansambel dengan pohon dengan menggunakan sampel pembelajaran asli lengkap beberapa kali. Prediksi untuk regresi dan klasifikasi dikumpulkan melalui suara mayoritas atau rata-rata aritmatika. Metode ini bertujuan untuk mengurangi varians dengan mengacak titik potong dan atribut secara eksplisit. Bias diminimalkan

dengan menggunakan sampel pembelajaran lengkap. Meskipun prosedur pemisahan node rumit, itu meningkatkan efisiensi komputasi.

Parameter dan masing-masing mempengaruhi pemilihan atribut, rata-rata kebisingan, dan pengurangan varians. Meskipun dapat diubah, pengaturan default lebih baik untuk otonomi metode dan komputasi. Algoritma Extra Trees Classifier digunakan dalam berbagai aplikasi seperti klasifikasi, prediksi, dan analisis data dan telah menjadi salah satu alat paling penting dalam model analisis data.

Prediksi dilakukan dalam klasifikasi dengan suara terbanyak dari pohon keputusan. Dengan kemampuannya yang kuat dalam memecahkan masalah klasifikasi dan prediksi, sangat berguna dalam pemodelan analisis data dan machine learning. Pengklasifikasi ekstra-pohon telah dipilih karena makna eksplisitnya, sifat sederhana, dan konversi yang mudah ke aturan "jika-maka". Metode extra-tree telah dipilih karena properti pengacakan untuk input numerik. Ide ini sangat berguna dalam masalah yang melibatkan sejumlah besar fitur numerik. Ini sering mengarah pada peningkatan akurasi.

2.6. Evaluasi

Model klasifikasi yang telah diuji selanjutnya dievaluasi untuk mengetahui performanya. Dalam penelitian ini, masalah yang dipecahkan adalah masalah klasifikasi, sehingga evaluasi yang tepat menggunakan matriks konfusi [36] Matriks konfusi adalah alat yang berguna untuk mengevaluasi kinerja model klasifikasi, terutama dalam situasi di mana kelas positif dan negatif memiliki proporsi yang tidak seimbang. Dengan informasi yang diberikan oleh matriks ini, kita dapat memahami di mana kesalahan model terjadi dan memutuskan tindakan yang tepat untuk meningkatkan kinerja model [37].

Pada matriks konfusi sendiri terdapat empat bagian, yaitu True Positive (TP) yang merupakan jumlah kasus di mana model dengan benar memprediksi kelas positif; False Positive (FP) adalah jumlah kasus di mana model salah memprediksi kelas positif padahal sebenarnya kelas negatif; True Negative (TN) adalah jumlah kasus di mana model dengan benar memprediksi kelas negatif; dan False Negative (FN) adalah jumlah kasus di mana model salah memprediksi kelas negatif padahal sebenarnya kelas positif. Beberapa indikator pengukuran digunakan dalam evaluasi, seperti akurasi (tingkat kebenaran model), sensitivitas / recall (rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif), presisi (rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif) dan F1-score (pembagian presisi dan sensitivitas. Berikut adalah rumus-rumusnya (Nengsih, 2019; Ghosh et al., 2021; Kurniawan dan Mustikasari, 2021).

$$Akurasi = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

$$Sensitivitas = \frac{TP}{TP+FN} \quad (6)$$

$$Presisi = \frac{TP}{TP+FP} \quad (7)$$

$$F1 - score = \frac{2*(precision* recall)}{(precision+recall)} \quad (8)$$

Akurasi berdasarkan rumus 5 mengukur seberapa sering model benar dalam memprediksi data secara keseluruhan, namun bisa menyesatkan jika data tidak seimbang. Untuk itu, sensitivitas (recall) berdasarkan rumus 6 fokus pada seberapa baik model mengenali data positif, tetapi bisa menghasilkan banyak prediksi positif yang salah. Presisi berdasarkan rumus 7 mengukur kualitas prediksi positif, yaitu seberapa banyak prediksi positif yang benar, namun sering kali berhadapan dengan trade-off dengan sensitivitas.

F1-score berdasarkan rumus 8 menggabungkan presisi dan sensitivitas dalam satu nilai, memberikan keseimbangan yang lebih adil dan lebih akurat dalam mengevaluasi model, terutama pada

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

dataset yang tidak seimbang. F1-score menjadi pilihan yang lebih baik karena mempertimbangkan keduanya secara seimbang.

3. HASIL DAN PEMBAHASAN

3.1. Preprocessing

Dataset yang digunakan memiliki empat baris data kosong atau nilai yang tidak ada pada fitur Albumin_and_Globulin_Ratio. Nilai-nilai ini kemudian diisi dengan nilai yang paling sering muncul pada kolom fitur atau. Dataset memiliki rentang nilai yang beragam antar fitur selain tidak memiliki nilai. Dataset dinormalisasi agar rentang nilai antar fiturnya seragam untuk memudahkan proses data mining.

Normalisasi data dilakukan menggunakan min- max karena metode ini memiliki kemampuan untuk mengatasi kelemahan metode seperti z-score atau penskalaan desimal. Berikutnya, data fitur gender dengan nilai "Female" diubah menjadi angka 0 dan nilai "Male" diubah menjadi angka 1 dapat diihat pada tabel 2.

Tabel 2. Normalisasi data

Fitur	Nilai Awal	Hasil Label Encoding
Gender	Female	0
	Male	1

Dimana ciri-ciri jenis kelamin yang nilai awalnya perempuan dan laki-laki adalah diubah ke bilangan biner yaitu 0 dan 1. Terakhir, label kelas pada kumpulan data yang memilki angka 1 dan 2 diubah menjadi angka 1 untuk kelas=1 dan 0 untuk kelas=2.

3.2. Proses SMOTE

Data kelas 1 lebih banyak dari data kelas 0 sehingga proses oversampling digunakan untuk menyeimbangkan persebaran data dengan SMOTE. Setelah SMOTE diterapkan, sebaran data dari fitur menjadi seimbang. Oversampling SMOTE berlaku untuk semua fitur dalam dataset, bukan hanya fitur Total_Billirubin. Maka didapatkan dataset dengan 416 baris data kelas 1 (positif) dan 0 (negatif), sehingga jumlah data di seluruh data menjadi seimbang, seperti yang ditunjukkan pada Tabel 3.

Tabel 3. Hasil Data Sebelum dan Sesudah SMOTE Berdasarkan Selector

Selector	Data Training	Data Testing
1	416	416
2	167	416

Dapat disimpulkan bahwa terdapat kemungkinan bahwa ketidakseimbangan data terjadi pada targetnya. Oleh karena itu, tahap selanjutnya adalah menghitung jarak instance bersama dengan jarak geometri pada target, dengan total 416 data. Untuk klasifikasi, dataset hasil SMOTE digunakan setelah penyeimbangan kelas selesai. Jumlah data dari masing-masing data seperti yang ditunjukkan pada tabel 4.

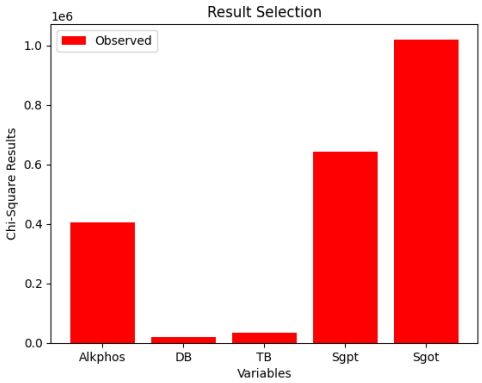
Tabel 4. Jumlah Data Uji dan Data Latih Keseluruhan

Kelas	Data Training	Data Testing
0	332	84
1	333	83

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

3.3. *Feature Selection*

Seleksi fitur dilakukan untuk menentukan fitur-fitur yang ada paling relevan dengan label untuk menyempurnakan model efisiensi dan efektivitas. Berikut ini adalah hasil perhitungan chi-square dari masing-masing fitur.

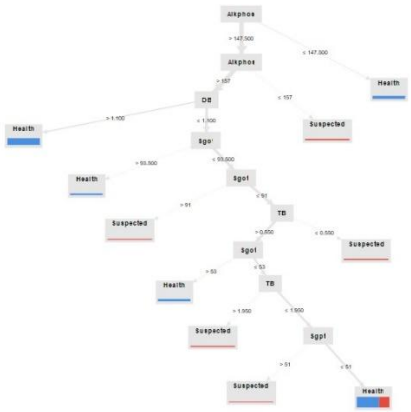


Gambar 2. Hasil Seleksi Fitur Chi-Scuare

Gambar 2 menunjukkan hasil seleksi fitur menggunakan metode chi-square pada dataset yang telah melalui proses SMOTE untuk mengatasi ketidakseimbangan data. Hasil analisis chi-square mengindikasikan bahwa fitur Sgot memiliki nilai tertinggi sebesar 1.0, menunjukkan ketergantungan yang kuat terhadap label target (penyakit liver). Fitur lainnya seperti Sgpt dengan nilai 0.62, Alkphos 0.4, TB 0.12, dan DB 0.11, juga memiliki nilai yang signifikan namun lebih rendah dibandingkan Sgot. Nilai chi-square yang tinggi menunjukkan perbedaan besar antara frekuensi yang diharapkan dan aktual, menandakan ketergantungan yang signifikan antara fitur dan label. Oleh karena itu, fitur-fitur dengan nilai chi-square tertinggi seperti Sgot, Sgpt, Alkphos, TB, dan DB dipilih untuk digunakan dalam pemodelan karena memiliki relevansi tinggi dalam memprediksi kondisi liver. Pemilihan ini didasarkan pada analisis statistik chi-square yang mengukur kekuatan asosiasi antara setiap fitur dengan variabel target, memastikan bahwa fitur-fitur tersebut berperan signifikan dalam klasifikasi penyakit liver.

3.4. Klasifikasi dengan Algoritma *Extra Trees Classifier*

Dengan menggunakan algoritma klasifikasi seperti Extra Trees Classifier, proses pengelompokan data menjadi lebih efisien dan dapat menghasilkan model prediktif yang memiliki akurasi tinggi serta kemampuan generalisasi yang baik terhadap data baru. Proses ini berlanjut hingga mencapai batas iterasi yang ditentukan. Berikut hasil pohon keputusan dari Extra Trees Classifier dan Chi-square



Gambar 3. Model Extra Trees Classifier

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

Gambar 3 menunjukkan dua warna: merah, nilai minimum, dan biru, nilai maksimum distribusi kumpulan data. Seperti yang telah kita ketahui, Extra Trees Classifier menghasilkan aturan-aturan yang dapat digunakan untuk mengambil keputusan dari data masukan. Berikut penjelasan mengenai peraturan yang ada. Tujuan dari analisis ini adalah untuk menggunakan algoritma Extra Trees Classifier dalam mengklasifikasikan data medis berdasarkan beberapa variabel. Variabel-variabel yang digunakan adalah Alkphos, DB, TB, Sgpt, dan Sgot.

3.5. Evaluasi

Pengukuran performa dari model yang telah dibuat menggunakan confusion matrix dimana hasil pengujian dapat dilihat pada tabel 5.

Tabel 5. Confusion Matrix

		Predicted	
		Positif (1)	Negatif (0)
Actual	Positif (1)	65	18
	Negatif (0)	11	73

Pada tabel di atas dapat dihitung akurasi hingga nilai F1-Score, berikut adalah perhitungan masing-masing indikator.

$$Akurasi = \frac{(65+73)}{(65+11+73+18)} = 0{,}826 \quad (9)$$

$$Sensitivitas = \frac{65}{65+11} = 0{,}855 \quad (10)$$

$$Presisi = \frac{65}{65+18} = 0{,}783 \quad (11)$$

$$F1 - score = \frac{2*(0{,}783*0{,}855)}{(0{,}783+0{,}855)} = 0{,}817 \quad (12)$$

Berdasarkan hasil perhitungan di atas, akurasi yang didapatkan adalah sebesar 0.826, yang menunjukkan bahwa model mampu mengklasifikasikan data dengan benar sekitar 82.6% dari total kasus. Sensitivitasnya sebesar 0.855, mengindikasikan model berhasil mengidentifikasi 85.5% dari kasus positif dengan benar. Presisi model tercatat sebesar 0.783, artinya dari semua prediksi positif, 78.3% adalah benar-benar positif. F1-Score yang diperoleh adalah 0.817, menunjukkan keseimbangan yang baik antara presisi dan sensitivitas. Secara keseluruhan, model Extreme Trees Classifier menunjukkan kinerja yang solid dan efektif dalam mengklasifikasikan data penyakit liver.

4. DISKUSI

Extra Trees Classifier dengan pemilihan fitur Chi-Square pada data pasien penyakit liver telah berhasil diterapkan pada penelitian ini. Pendekatan inovatif ini telah menyederhanakan proses diagnostik dan meningkatkan keakuratannya secara signifikan. Dengan menganalisis dan memproses kumpulan data dari Kaggle secara cermat, metodologi ini telah menetapkan standar baru dalam penggunaan pembelajaran mesin dalam diagnostik medis, khususnya dalam deteksi dini penyakit liver, sebuah langkah penting dalam meningkatkan harapan hidup pasien.

Di antara segudang fitur yang tersedia dalam dataset, lima dipilih sebagai yang paling berpengaruh melalui proses pemilihan fitur Chi-Square: Sgot, Sgpt, Alkphos, TB, dan DB. Pemilihan lima fitur ini terbukti lebih efektif dibandingkan penelitian sebelumnya yang menggunakan lebih banyak fitur. Sebagai contoh, penelitian Yang et al. (2023) menggunakan 10 fitur namun hanya mencapai

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

akurasi 78,5%. Pengurangan jumlah fitur tidak hanya meningkatkan efisiensi komputasi tetapi juga mengurangi risiko overfitting. Ciri-ciri ini sangat penting karena berkaitan langsung dengan fungsi dan kesehatan hati, menjadikannya penanda yang sangat diperlukan untuk penyakit liver. Pemilihan mereka menggarisbawahi kemampuan model untuk fokus pada variabel yang relevan secara klinis, sehingga meningkatkan penerapan dan keandalannya. Dalam mengimplementasikan model, kita dapat melihat penelitian terkait sebelumnya.

Tabel 6. Penelitian Terkait

Studi Sebelumnya	Metode yang Digunakan	Hasil
[39]	SMOTE-Support Vector Machine	65%
[40]	SMOTE-XGBoost-Bayesian Search	76.7 %
[41]	SMOTE-Random Forest	77,06%
[42]	ADASYN- Support Vector Machine	80,4%
Metode yang Diusulkan	SMOTE-Chi square-Extra Trees Classifier	82,6%

Temuan penelitian ini, seperti yang ditunjukkan pada Tabel 6, menunjukkan tingkat akurasi yang tertinggi sebesar 82,6% melampaui upaya penelitian sebelumnya dalam domain ini. Akurasi yang unggul ini merupakan bukti kemanjuran penggabungan Extra Trees Classifier dengan pemilihan fitur Chi-Square dalam meningkatkan kemampuan prediktif model machine learning untuk deteksi dini penyakit liver. Kemajuan tersebut berkontribusi pada bidang akademis dengan menyediakan model yang kuat untuk prediksi penyakit dan menawarkan implikasi praktis bagi para profesional kesehatan. Dengan mengadopsi model ini, mereka dapat mencapai diagnosis yang lebih akurat secara dini, sehingga meningkatkan hasil pasien melalui intervensi yang tepat waktu dan tepat sasaran. Oleh karena itu, penelitian ini tidak hanya menandai langkah maju yang signifikan dalam penerapan pembelajaran mesin dalam layanan kesehatan tetapi juga membuka landasan bagi inovasi masa depan dalam diagnosis dan pengobatan penyakit liver.

Meskipun penelitian ini menawarkan pendekatan inovatif dengan akurasi yang mengesankan dalam deteksi dini penyakit liver melalui integrasi Extra Trees Classifier dengan pemilihan fitur Chi-Square, terdapat keterbatasan, seperti ketergantungan pada kualitas kumpulan data dari Kaggle, yang mungkin tidak dapat diandalkan. mencakup demografi pasien yang luas atau data yang tidak lengkap, serta potensi pengabaian interaksi antar variabel yang dapat memberikan wawasan tambahan. Namun, implikasinya sangat signifikan, yaitu menawarkan metode diagnostik yang lebih efisien dan akurat untuk deteksi dini penyakit liver, yang dapat meningkatkan harapan hidup pasien melalui intervensi yang tepat waktu dan tepat sasaran serta mendorong inovasi lebih lanjut dalam penerapan machine learning dalam diagnostik medis dan perawatan kesehatan.

5. KESIMPULAN

Liver atau hati adalah organ vital yang berfungsi menetralisir racun, mengatur sirkulasi hormon, dan membantu pencernaan lemak. Penyakit liver, sering disebut sebagai pembunuh diam karena tidak menunjukkan gejala awal, dapat disebabkan oleh pola hidup tidak sehat, infeksi, kelainan bawaan, kecanduan alkohol, dan merokok. Keterlambatan diagnosis seringkali menyebabkan kondisi kesehatan pasien memburuk, sehingga diperlukan pemeriksaan rutin untuk deteksi dini dan pengobatan yang tepat.

Penelitian ini menggunakan algoritma machine learning, Extra Trees Classifier, dan seleksi fitur chi-square untuk mendeteksi penyakit liver sejak dini. Hasilnya menunjukkan akurasi model sebesar 82,6%, sensitivitas 85,5%, dan presisi 78,3%. Metodologi ini meningkatkan akurasi diagnostik dengan memilih fitur paling berpengaruh, yaitu Sgot, Sgpt, Alkphos, TB, dan DB. Penggabungan Extra Trees

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

Classifier dengan seleksi fitur chi-square terbukti lebih akurat dibandingkan metode sebelumnya, seperti SMOTE-Support Vector Machine.

Kesimpulan penelitian ini menegaskan bahwa pendekatan yang digunakan mampu meningkatkan akurasi diagnostik penyakit liver. Dengan akurasi sebesar 82,6% dan sensitivitas 85,5%, metode ini menunjukkan keunggulan dalam deteksi dini. Urgensi penelitian ini terletak pada meningkatnya prevalensi penyakit liver dan keterbatasan metode diagnostik konvensional. Dengan model yang lebih akurat, deteksi dini dapat ditingkatkan sehingga memungkinkan intervensi medis lebih cepat dan efektif.

Ke depannya, penelitian diharapkan dapat mengembangkan model yang lebih kompleks, mengintegrasikan data multimodal, dan meningkatkan kualitas data. Selain itu, pengembangan alat diagnostik real-time, studi longitudinal, dan kolaborasi interdisipliner juga menjadi fokus. Upaya ini bertujuan untuk menemukan metode deteksi dini yang lebih efisien dan akurat, meningkatkan harapan hidup pasien, serta mendorong inovasi dalam penerapan machine learning dalam diagnostik medis dan perawatan kesehatan.

UCAPAN TERIMA KASIH

Ucapan terima kasih dapat diberikan setelah kesimpulan dan sebelum daftar pustaka. Penulis dapat menuliskan bagian ini ataupun menghapusnya. Ucapan terima kasih hanya diperuntukkan bagi penyandang dana dan objek penelitian saja. Penulisan ucapan terima kasih diluar 2 hal tersebut tidak diperbolehkan.

DAFTAR PUSTAKA

[1] E. Patimah, V. B. Haekal, and D. Sandya Prasvita, "Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision Tree," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, vol. 2, no. 1, pp. 655–659, 2021.

[2] B. N. P. Maharani, A. D. Hendriani, and P. W. P. Iswari, "Liver Cirrhosis: Pathophysiology, Diagnosis, and Management," *J. Biol. Trop.*, vol. 23, no. 1, pp. 457–463, 2023, doi: 10.29303/jbt.v23i1.5763.

[3] M. Ghosh *et al.*, "A comparative analysis of machine learning algorithms to predict liver disease," *Intell. Autom. Soft Comput.*, vol. 30, no. 3, pp. 917–928, 2021, doi: 10.32604/iasc.2021.017989.

[4] Z. Guo *et al.*, "A randomized-controlled trial of ischemia-free liver transplantation for end-stage liver disease," *J. Hepatol.*, vol. 79, no. 2, pp. 394–402, 2023, doi: 10.1016/j.jhep.2023.04.010.

[5] M. E. Rinella *et al.*, *AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease*, vol. 77, no. 5. 2023. doi: 10.1097/HEP.0000000000000323.

[6] L. Rong, J. Zou, W. Ran, and X. Qi, "Advancements in the treatment of non-alcoholic fatty liver disease (NAFLD)," no. January, pp. 1–18, 2023, doi: 10.3389/fendo.2022.1087260.

[7] M. V. Machado, "Aerobic exercise in the management of metabolic dysfunction associated fatty liver disease," 2021. doi: 10.2147/DMSO.S304357.

[8] A. S. Afrah, "Sistem Diagnosa Penyakit Liver Menggunakan Metode Artificial Neural Network: Studi Berdasarkan Dataset Indian Liver Patient Dataset," *J. Inform. J. Pengemb. IT*, vol. 8, no. 3, pp. 308–312, Dec. 2023, doi: 10.30591/jpit.v8i3.5346.

[9] M. T. Long, M. Noureddin, and J. K. Lim, "CLINICAL PRACTICE UPDATE AGA Clinical Practice Update : Diagnosis and Management Expert Review," *Gastroenterology*, vol. 163, no. 3, pp. 764-774.e1, 2022, doi: 10.1053/j.gastro.2022.06.023.

[10] D. S. Ali and M. A. Aljabery, "Predicting Liver Cirrhosis Stages Using Extra Trees, Random Forest, and SVM with Data Mining Techniques," *Inform.*, vol. 48, no. 21, pp. 15–26, 2024, doi: 10.31449/inf.v48i21.6752.

[11] F. Muhammad *et al.*, "Liver Ailment Prediction Using Random Forest Model," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1049–1067, 2023, doi: 10.32604/cmc.2023.032698.

[12] Y. O. Daddala and K. Shaik, "Cardiovascular Disease Prediction: Employing Extra Tree Classifier-Based Feature Selection and Optimized RNN with Artificial Bee Colony," *Rev.*

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

d'Intelligence Artif., vol. 38, no. 2, pp. 643–653, Apr. 2024, doi: 10.18280/ria.380228.
[13] Y. Duan *et al.*, "Association of Inflammatory Cytokines With Non-Alcoholic Fatty Liver Disease," *Front. Immunol.*, vol. 13, no. May, 2022, doi: 10.3389/fimmu.2022.880298.
[14] D. Sharma, R. Kumar, and A. Jain, "Measurement : Sensors Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning," *Meas. Sensors*, vol. 24, no. September, p. 100560, 2022, doi: 10.1016/j.measen.2022.100560.
[15] M. Mahmud *et al.*, "Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 116–124, Mar. 2024, doi: 10.35882/jeeemi.v6i2.384.
[16] 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.
[17] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach," *Chemom. Intell. Lab. Syst.*, vol. 222, no. 06, p. 104516, Mar. 2022, doi: 10.1016/j.chemolab.2022.104516.
[18] P. Theerthagiri, "Liver disease classification using histogram-based gradient boosting classification tree with feature selection algorithm," *Biomed. Signal Process. Control*, vol. 100, p. 107102, Feb. 2025, doi: 10.1016/j.bspc.2024.107102.
[19] A. Panwar, V. Bhatnagar, M. Khari, A. W. Salehi, and G. Gupta, "A Blockchain Framework to Secure Personal Health Record (PHR) in IBM Cloud-Based Data Lake," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3045107.
[20] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study," *J. Diabetes Metab. Disord.*, vol. 21, no. 1, pp. 251–261, Jan. 2022, doi: 10.1007/s40200-021-00968-z.
[21] S. Qin *et al.*, "Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults," *Sci. Rep.*, vol. 13, no. 1, pp. 1–7, 2023, doi: 10.1038/s41598-023-30750-5.
[22] K. Stefanus and H. Leong, "Comparison of Random Forest Algorithm Accuracy With Xgboost Using Hyperparameters," *Proxies J. Inform.*, vol. 7, no. 1, pp. 15–23, 2024, doi: 10.24167/proxies.v7i1.12464.
[23] D. Baby, S. J. Devaraj, J. Hemanth, and M. M. Anishin Raj, "Leukocyte classification based on feature selection using extra trees classifier: A transfer learning approach," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 29, no. 8, pp. 2742–2757, 2021, doi: 10.3906/elk-2104-183.
[24] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, 2023, doi: 10.3390/biomedicines11020581.
[25] F. ORHANBULUCU, A. İrem, F. LATİFOĞLU, and İ. Semra, "Predicting liver disease using decision tree ensemble methods," *Erciyes Üniversitesi Fen Bilim. Enstitüsü Fen Bilim. Derg.*, vol. 38, no. 2, pp. 261–267, 2022.
[26] K. R. Makkena and K. Natarajan, "Classification Algorithms for Liver Epidemic Identification," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 9, pp. 1–13, 2023, doi: 10.4108/eetpht.9.4379.
[27] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.
[28] A. A. Kurniawan and M. Mustikasari, "Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 544, 2021, doi: 10.32493/informatika.v5i4.6760.
[29] R. Atiq, F. Fariha, M. Mahmud, S. S. Yeamin, K. I. Rushee, and S. Rahim, "A Comparison of Missing Value Imputation Techniques on Coupon Acceptance Prediction," *Int. J. Inf. Technol. Comput. Sci.*, vol. 14, no. 5, pp. 15–25, 2022, doi: 10.5815/ijitcs.2022.05.02.
[30] I. Huda, "Implementasi Natural Language Processing (Nlp) Untuk Aplikasi Pencarian Lokasi," *J. Nas. Teknol. Terap.*, vol. 3, no. 2, p. 15, 2021, doi: 10.22146/jntt.35036.
[31] A. Hristov, A. Tahchiev, H. Papazov, N. Tulechki, T. Primov, and S. Boytcheva, "Application

Vol. 6, No. 5, October 2025, Page. 3405-3418
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4261

of Deep Learning Methods to SNOMED CT Encoding of Clinical Texts: From Data Collection to Extreme Multi-Label Text-Based Classification," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2021, pp. 557–565. doi: 10.26615/978-954-452-072-4_063.

[32] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over - sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Inform. Decis. Mak.*, vol. 2, pp. 1–14, 2022, doi: 10.1186/s12911-022-02075-2.

[33] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods," *Expert Syst. Appl.*, vol. 178, no. April, p. 114986, Sep. 2021, doi: 10.1016/j.eswa.2021.114986.

[34] A. R. B. Alamsyah, S. R. Anisa, N. S. Belinda, and A. Setiawan, "SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1, pp. 305–314, 2022, doi: 10.34123/icdsos.v2021i1.240.

[35] H. M. Qasim, O. Ata, M. A. Ansari, M. N. Alomary, S. Alghamdi, and M. Almehmadi, "Hybrid Feature Selection Framework for the Parkinson Imbalanced Dataset Prediction Problem," *Medicina (B. Aires).*, vol. 57, no. 11, p. 1217, Nov. 2021, doi: 10.3390/medicina57111217.

[36] D. Salirawati, "Identifikasi Problematika Evaluasi Pendidikan Karakter di Sekolah," *J. Sains dan Edukasi Sains*, vol. 4, no. 1, pp. 17–27, 2021, doi: 10.24246/juses.v4i1p17-27.

[37] K. M. Elistiana, B. A. Kusuma, P. Subarkah, and H. A. A. Rozaq, "Improvement of Naive Bayes Algorithm in Sentiment Analysis of Shopee Application Reviews on Google Play Store," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1431–1436, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1486.

[38] W. Nengsih, "Analisa Akurasi Permodelan Supervised Dan Unsupervised Learning Menggunakan Data Mining," *Sebatik*, vol. 23, no. 2, pp. 285–291, 2019, doi: 10.46984/sebatik.v23i2.771.

[39] P. J. Shetty, "Prediction performance of classification models for imbalanced liver disease data," vol. 8, no. 5, pp. 58–62, 2023.

[40] R. Ubaidillah, M. Muliadi, D. T. Nugrahadi, M. R. Faisal, and R. Herteno, "Implementasi XGBoost Pada Keseimbangan Liver Patient Dataset dengan SMOTE dan Hyperparameter Tuning Bayesian Search," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 3, p. 1723, Jul. 2022, doi: 10.30865/mib.v6i3.4146.

[41] M. A. Khadija and N. A. Setiawan, "Detecting Liver Disease Diagnosis by Combining SMOTE, Information Gain Attribute Evaluation and Ranker," *ITSMART J. Teknol. dan Inf.*, vol. 9, no. 1, pp. 13–17, 2020.

[42] W. Hidayat, M. Ardiansyah, and A. Setyanto, "Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb," *Edumatic J. Pendidik. Inform.*, vol. 5, no. 1, pp. 11–20, 2021, doi: 10.29408/edumatic.v5i1.3125.