DECODE: JURNAL PENDIDIKAN TEKNOLOGI INFORMASI
P-ISSN: 2775-2984 | E-ISSN: 2775-1813
Vol. 6, No 1, Maret 2026, Hal: 223-236
Doi: http://dx.doi.org/10.51454/decode.v6i1.1630
https://journal.umkendari.ac.id/index.php/decode

Analisis Klasifikasi Hepatitis Menggunakan Synthetic Minority Oversampling Technique, *Support Vector Machine*, dan Random Forest

Amalia Nur Laily[1*], Mula Agung Barata[1], Denny Nurdiansyah[2]

[1]Program Studi Teknik Informatika, Universitas Nahdlatul Ulama Sunan Giri, Indonesia
[2]Program Studi Statistika, Universitas Nahdlatul Ulama Sunan Giri, Indonesia.

Artikel Info

Kata Kunci:

Hepatitis;
Random Forest;
SMOTE;
Support Vector Machine.

Keywords:

Hepatitis;
Random Forest;
SMOTE;
Support Vector Machine.

Riwayat Artikel:

Submitted: 10 Februari 2026
Accepted: 31 Maret 2026
Published: 31 Maret 2026

Abstrak: Hepatitis akibat infeksi virus masih menjadi masalah kesehatan masyarakat yang serius sehingga deteksi dini berbasis data klinis penting untuk mencegah kerusakan hati lebih lanjut. Penelitian ini menganalisis kinerja algoritma *Support Vector Machine* (SVM) dan *Random Forest* pada klasifikasi hepatitis serta mengkaji dampak penerapan *Synthetic Minority Over-sampling Technique* (*SMOTE*). Dataset yang digunakan adalah *HepatitisCdata.csv* dari *Kaggle* dengan 615 data pasien yang memuat atribut demografis dan parameter biokimia hati. Tahapan penelitian meliputi *preprocessing* data, penanganan *outlier*, transformasi atribut kategorikal, serta pembangunan model *baseline* dan *SMOTE*. Evaluasi dilakukan menggunakan 10-*fold cross-validation* dengan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil menunjukkan bahwa *SMOTE* meningkatkan performa kedua algoritma, dengan *Random Forest* + *SMOTE* memberikan hasil terbaik (akurasi 98,85%) dibandingkan SVM + *SMOTE* (98,50%). Kontribusi penelitian ini terletak pada penggunaan pipeline *preprocessing* dan evaluasi yang seragam untuk membandingkan dampak *SMOTE* secara langsung pada dua algoritma klasifikasi hepatitis.

Abstract: *Hepatitis caused by viral infection remains a serious public health problem, making early detection based on clinical data important to prevent further liver damage. This study analyzes the performance of Support Vector Machine (SVM) and Random Forest algorithms for hepatitis classification and examines the impact of applying Synthetic Minority Over-sampling Technique (SMOTE). The dataset used was HepatitisCdata.csv from Kaggle, consisting of 615 patient records containing demographic attributes and liver biochemical parameters. The research stages included data preprocessing, outlier handling, categorical attribute transformation, and the development of baseline and SMOTE-based models. Evaluation was conducted using 10-fold cross-validation with accuracy, precision, recall, and F1-score metrics. The results showed that SMOTE improved the performance of both algorithms, with Random Forest + SMOTE achieving the best result (98.85% accuracy) compared with SVM + SMOTE (98.50%). The contribution of this study lies in the use of a consistent preprocessing and evaluation pipeline to directly compare the impact of SMOTE on two hepatitis classification algorithms.*

Corresponding Author:
Amalia Nur Laily
Teknik Informatika, Fakultas Sains Dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri
Alamat: Jl. Jendral Ahmad Yani No.10, Jambean, Sukorejo, Kec. Bojonegoro, Kabupaten Bojonegoro, Jawa Timur 62115, Indonesia
Email: amalianurlaily2004@gmail.com

PENDAHULUAN

Infeksi virus hepatitis, khususnya hepatitis B dan C, merupakan salah satu penyebab utama kematian akibat penyakit menular di dunia dan masih menjadi tantangan besar bagi sistem kesehatan *global*. Berdasarkan *Global Hepatitis Report 2024*, diperkirakan sekitar 1,3 juta orang meninggal akibat hepatitis virus pada tahun 2022, termasuk hampir 1,1 juta kematian yang berkaitan dengan hepatitis B dan sekitar 244.000 kematian akibat hepatitis C (Zhang & Cui, 2025). Angka tersebut menunjukkan bahwa target WHO untuk menurunkan infeksi baru sebesar 90% dan mortalitas sebesar 65% pada tahun 2030 masih jauh dari tercapai, terutama di kawasan Afrika, Asia Tenggara, dan Pasifik Barat yang menanggung proporsi terbesar kasus dan kematian. Sebagian besar beban ini sebenarnya dapat dikurangi melalui perluasan cakupan vaksinasi, peningkatan skrining laboratorium, serta akses pengobatan antivirus yang terjangkau, yang didukung oleh sistem deteksi dini berbasis data klinis yang terintegrasi dan tepat waktu.

Kondisi serupa juga terlihat di Indonesia, di mana berbagai laporan epidemiologi nasional dan regional menunjukkan bahwa infeksi hepatitis B dan C tetap menjadi penyebab utama sirosis dan karsinoma hepatoseluler, dengan beban kasus yang signifikan pada kelompok usia produktif. Studi di fasilitas pelayanan kesehatan dan rumah sakit rujukan menunjukkan bahwa banyak pasien hepatitis datang pada fase lanjut, dengan gangguan fungsi hati yang telah jelas pada parameter biokimia seperti enzim hati, *bil*irubin, dan penanda infeksi virus (Kementrian Kesehatan, 2016) (Kementerian Kesehatan RI, 2018). Hal ini mengindikasikan bahwa deteksi dini berbasis pemeriksaan rutin masih terbatas. Dalam konteks tersebut, pemanfaatan teknik *machine learning* dan *data mining* pada data klinis serta hasil pemeriksaan laboratorium pasien hepatitis muncul sebagai strategi potensial untuk meningkatkan efisiensi skrining, klasifikasi tingkat keparahan penyakit, dan akurasi diagnosis secara sistematis dan konsisten (Alnur et al., 2023).

Salah satu algoritma yang banyak digunakan dalam klasifikasi data medis, termasuk penyakit hepatitis, adalah *Support Vector Machine* (*SVM*), yang bekerja dengan mencari hiperbidang pemisah bermargin maksimum sehingga efektif membedakan kelas pasien berdasarkan kombinasi fitur klinis dan laboratorik yang kompleks. Di sisi lain, *Random Forest* memanfaatkan kumpulan pohon keputusan yang dibangun dari subset data dan fitur yang berbeda, sehingga menghasilkan model yang relatif sta*bil* terhadap *noise* dan mampu menangkap hubungan nonlinier antar fitur. Studi sebelumnya menunjukkan bahwa *SVM* dan *Random Forest* mampu mengklasifikasikan kasus hepatitis dengan akurasi tinggi, dengan *SVM* ber-kernel Gaussian RBF memberikan kinerja terbaik di antara konfigurasi yang diuji (Aurelia et al., 2021). Dukungan empiris terhadap sta*bil*itas *SVM* pada data tabular berdimensi tinggi juga ditunjukkan pada penelitian lain yang melaporkan bahwa *SVM* mampu mencapai kinerja klasifikasi yang sangat tinggi setelah melalui tahapan *preprocessing* dan seleksi fitur yang sistematis pada data dunia nyata yang heterogen, sehingga memperkuat relevansi penggunaan *SVM* pada data klinis hepatitis yang kompleks dan multidimensi (Khairunnas et al., 2025).

Pada skala data yang lebih besar, perbandingan beberapa algoritma pembelajaran mesin untuk prediksi hepatitis C menunjukkan bahwa *Random Forest* mencapai akurasi tertinggi serta tingkat kesalahan klasifikasi terendah (Gunawan & Ilham Pratama, 2024). Evaluasi komprehensif terhadap beberapa model pada dataset hepatitis UCI juga menemukan bahwa *Random Forest* dengan seleksi fitur Boruta menghasilkan *F1-score* dan sensitivitas tertinggi dibandingkan *SVM* dan model lainnya (Khatun et al., 2025).Dalam konteks ketidakseimbangan kelas, penerapan *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada klasifikasi hepatitis C berbasis *Random Forest* terbukti meningkatkan akurasi serta kualitas pemisahan antara kelas terinfeksi dan tidak terinfeksi (Sharfina & Ramadhan, 2023). Kombinasi *SMOTE* dan seleksi fitur berbasis *ensemble filter* pada *SVM* juga terbukti mampu memperbaiki performa klasifikasi penyakit liver, khususnya pada kelas minoritas (Nugraha et al., 2023). Temuan tersebut diperkuat oleh studi lain yang menunjukkan bahwa pada data dengan ketidakseimbangan kelas ekstrem, penerapan *SMOTE* pada *Random Forest* mampu meningkatkan *recall* kelas minoritas secara drastis, dari kondisi gagal terdeteksi menjadi hampir sempurna, sehingga menegaskan pentingnya penanganan ketidakseimbangan kelas secara eksplisit (Putri & Rachmatika, 2025).

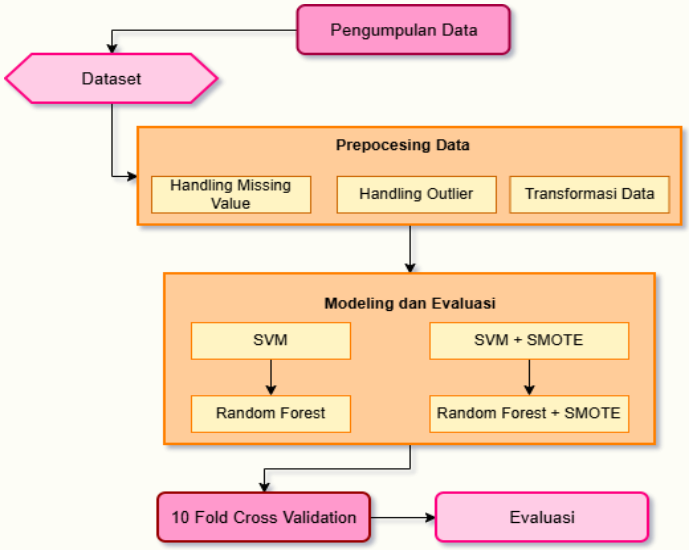
Pendekatan lanjutan dalam pengembangan model klasifikasi hepatitis juga menekankan pentingnya rancangan *pipeline* yang sistematis. Kerangka *explainable machine learning* yang mengombinasikan beberapa algoritma klasifikasi dengan seleksi fitur *sequential forward selection* terbukti mampu meningkatkan akurasi sekaligus mempertahankan interpreta*bil*itas model (Ali et al., 2023). Integrasi seleksi fitur dan penalaan parameter pada *Random Forest* juga dilaporkan mampu menghasilkan model dengan jumlah fitur lebih sedikit namun akurasi dan sta*bil*itas validasi yang lebih tinggi (Farghaly et al., 2023).Kerangka prediksi hepatitis C berbasis berbagai model pembelajaran mesin menunjukkan bahwa kombinasi seleksi fitur dan *SMOTE* pada *Random Forest* menghasilkan akurasi, *Brier score*, dan *Matthews correlation coefficient* terbaik (Kumari et al., 2025). Kinerja prediktif unggul dari algoritma berbasis pohon keputusan dan *gradient boosting* juga dilaporkan pada data hepatitis C (Cabanillas-Carbonell & Zapata-Paulini, 2025), sementara penerapan *SMOTE* terbukti meningkatkan kemampuan *Random Forest* dan XGBoost dalam mengenali tingkat fibrosis dan sirosis yang sebelumnya sulit terdeteksi (Syukron et al., 2020).Pentingnya integrasi seleksi fitur, penanganan ketidakseimbangan data, dan evaluasi komparatif algoritma juga ditekankan dalam diagnosis dini penyakit hati (Rehman et al., 2024). Pada konteks hepatitis C, integrasi *SMOTE*, optimasi hiperparameter berbasis Optuna, serta analisis interpreta*bil*itas menggunakan SHAP terbukti meningkatkan performa deteksi infeksi HCV secara signifikan (Mehzabeen et al., 2025). Pendekatan hibrida melalui penggabungan *Random Forest* dan *SVM* juga menghasilkan model yang lebih presisi dan tahan terhadap *overfitting* (Lilhore et al., 2023).

Pada klasifikasi multi-kelas hepatitis, *SVM* dengan kernel linier dilaporkan mencapai kinerja tertinggi dan *F1-score* yang seimbang (Rehman et al., 2024). Meskipun *Logistic Regression* menunjukkan performa yang kompetitif, *SVM* tetap unggul dalam hal akurasi dan AUC (Alnur et al., 2023). Selain itu, penerapan *SMOTE* pada *Random Forest* secara konsisten meningkatkan sensitivitas dan *F1-score* pada data kesehatan yang tidak seimbang (Erlin et al., 2022). Secara keseluruhan, temuan-temuan tersebut memperkuat bahwa kombinasi *SVM* dan *Random Forest* yang dirancang dalam *pipeline* dengan seleksi fitur, *SMOTE*, dan optimasi hiperparameter merupakan fondasi yang kuat untuk pengembangan sistem klasifikasi hepatitis. Hal ini sejalan dengan temuan sebelumnya yang menegaskan efektivitas pendekatan *data mining* dan model ansambel dalam mendukung pengam*bil*an keputusan berbasis data klinis (Barata et al., 2021; Barata et al., 2025; Purnomo et al., 2020; Yaqin et al., 2025).

Berdasarkan tinjauan tersebut, dapat disimpulkan bahwa meskipun *SVM* dan *Random Forest* telah terbukti efektif dalam klasifikasi hepatitis, sebagian besar penelitian sebelumnya belum mengkaji secara sistematis dampak penerapan *SMOTE* dalam satu kerangka eksperimen yang terstandarisasi. Oleh karena itu, penelitian ini berfokus pada peningkatan kinerja klasifikasi hepatitis melalui penerapan *SMOTE* pada algoritma *SVM* dan *Random Forest* dalam *pipeline* pra-pemrosesan dan validasi silang yang eksplisit dan terkontrol.

METODE

Secara garis besar, alur penelitian ini terdiri atas beberapa tahap utama, yaitu: (1) pengumpulan data kasus hepatitis sebagai sumber data awal, (2) tahap praproses data yang meliputi penanganan nilai hilang (*missing* value) identifikasi penanganan *outlier* serta transformasi data , (3) penanganan ketidakseimbangan kelas pada data pelatihan menggunakan teknik *SMOTE*, (4) pembangunan model klasifikasi pada tahap data mining dengan menggunakan dua algoritma, yakni *Support Vector Machine* dan *Random Forest*, serta (5) pengujian dan pengukuran kinerja model melalui *skema 10-fold cross validation* hingga seluruh rangkaian eksperimen selesai dilaksanakan. Alur lengkap tahapan penelitian tersebut ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

Dataset

Dataset yang digunakan dalam penelitian ini bersumber dari platform *Kaggle*, dengan nama dataset "*HepatitisCdata.csv*". Dataset ini terdiri dari sejumlah atribut yang dikelompokkan menjadi variabel independen dan satu variabel dependen berupa label status hepatitis. Variabel independennya mencakup atribut demografis (seperti usia dan jenis kelamin) serta berbagai parameter klinis dan laboratorik yang berkaitan dengan fungsi hati, antara lain nilai enzim hati, kadar *bil*irubin, dan indikator biokimia lain yang lazim digunakan dalam evaluasi gangguan hati; sedangkan variabel dependennya adalah label status hepatitis yang menyatakan apakah pasien termasuk dalam kategori hepatitis (1) atau non-hepatitis (0). Pemilihan dataset ini sejalan dengan berbagai penelitian yang memanfaatkan data hepatitis klinis terstruktur dengan komposisi fitur demografis dan parameter fungsi hati yang serupa untuk pengembangan model klasifikasi berbasis machine learning pada penyakit hepatitis (Gunawan & Ilham Pratama, 2024; Kumari et al., 2025). Pada tahap ini juga dilakukan peninjauan awal terhadap struktur data untuk memastikan bahwa tipe data setiap atribut telah sesuai dengan kebutuhan analisis berikutnya. Rincian masing-masing fitur dapat dilihat pada Tabel 1.

Tabel 1. Dataset

No	Ctg	Age	Sex	Alb	Alp	Alt	Ast	Bil	Che	Chol	Crea	Ggt	Prot
0	0	32.0	0	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	0	32.0	0	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	0	32.0	0	46.9	74.7	36.2	49.9	6.1	8.84	5.20	86.0	33.2	79.3
3	0	32.0	0	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	0	32.0	0	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...
610	1	62.0	1	32.0	118.8	5.9	49.9	20.1	5.57	6.30	55.7	77.0	68.5
611	1	64.0	1	29.2	102.8	2.9	44.4	20.0	2.95	3.02	63.0	35.9	71.3
612	1	64.0	1	29.2	87.3	3.5	49.9	20.1	2.95	3.63	66.7	64.2	82.0
613	1	46.0	1	33.0	66.2	39.0	49.9	20.0	3.56	4.20	52.0	50.0	71.0
614	1	59.0	1	36.0	66.2	58.0	49.9	12.0	9.07	5.30	67.0	34.0	68.0

***Preprocessing* Data**

Setelah dataset hepatitis diperoleh, tahap berikutnya adalah *preprocessing* data. Pada tahap ini, fokus utama adalah meningkatkan kualitas data agar siap digunakan pada proses pemodelan. Pertama, dilakukan identifikasi dan penanganan nilai hilang (*missing value*) pada setiap atribut. Atribut yang mengandung nilai hilang dianalisis distribusinya, kemudian diterapkan teknik imputasi yang sesuai, misalnya imputasi menggunakan nilai tengah (*median*) atau pendekatan lain yang konsisten dengan

karakteristik data, sehingga kehilangan informasi dapat diminimalkan tanpa harus menghapus terlalu banyak sampel.

Secara umum, untuk suatu atribut numerik

$$X = \{x_1, x_2, \dots, x_n\} \quad (1)$$

Yang telah diurutkan sehingga

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}, \quad (2)$$

Median dapat didefinisikan sebagai:

jika n ganjil

$$median(X) = x_{\left(\frac{n+1}{2}\right)} \quad (3)$$

jika n genap

$$median(X) = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \quad (4)$$

Dimana:

X	: himpunan nilai suatu atribut numerik (misalnya kadar enzim hati).
$x_{(k)}$	: nilai ke-k setelah data diurutkan dari kecil ke besar.
$median(X)$	: nilai tengah distribusi yang digunakan untuk imputasi nilai hilang.

Kedua, dilakukan deteksi dan penanganan *outlier* pada atribut numerik, khususnya pada parameter laboratorium yang berpotensi memiliki nilai ekstrem. *Outlier* diidentifikasi dengan kriteria statistik tertentu, kemudian disesuaikan melalui pendekatan seperti *capping* atau transformasi nilai, dengan tetap mempertahankan kewajaran secara klinis. Hasil dari tahap ini adalah dataset yang lebih bersih, sta*bil*, dan representatif, sehingga mengurangi risiko bias serta gangguan terhadap proses pembelajaran model *Support Vector Machine* dan *Random Forest* pada tahap berikutnya.

Pada tahap ketiga, dilakukan transformasi data melalui verifikasi terhadap variabel target *Category*. Kolom ini telah dikonversi pada langkah praproses sebelumnya ke dalam bentuk numerik, sehingga pada tahap ini hanya dilakukan pengecekan nilai unik untuk memastikan bahwa representasi biner tersebut sudah konsisten dan tidak memerlukan konversi ulang. Transformasi ini memastikan bahwa variabel target dan atribut *Sex* berada dalam format numerik yang seragam, sehingga mendukung stabilitas proses pelatihan model *Support Vector Machine* dan *Random Forest* pada tahap *data mining*.

Smote

Dalam penelitian ini, perhatian khusus diberikan pada karakteristik ketidakseimbangan kelas (class imbalance) yang umum terjadi pada data medis, termasuk pada kasus hepatitis, di mana jumlah sampel pasien yang terdiagnosis hepatitis sering kali lebih sedikit dibandingkan pasien non-hepatitis. Ketimpangan tersebut dapat membuat model lebih condong mempelajari pola pada kelas mayoritas, sementara kelas minoritas berisiko kurang terwakili atau bahkan terabaikan dalam proses pembelajaran sehingga menurunkan kemampuan deteksi kasus *positif*. Untuk mengatasi hal tersebut, diterapkan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada data latih. *SMOTE* bekerja dengan membangkitkan sampel sintetis baru di sekitar contoh-contoh kelas minoritas dalam ruang fitur secara matematis, untuk suatu sampel minoritas x_i dan salah satu sampel lain pada kelas minoritas yang dijadikan acuan interpolasi, dilambangkan sebagai x_j sampel sintetis x_{baru} dihasilkan sebagai berikut:

$$x_{\text{baru}} = x_i + \lambda\left(x_j - x_i\right), \quad \lambda \sim U(0,1) \quad (5)$$

Dimana:

x_i	: vektor fitur salah satu sampel pada kelas minoritas (pasien hepatitis),
x_j	: vektor fitur sampel lain kelas minoritas sebagai titik acuan interpolasi *SMOTE*,
λ	:*bil*angan acak yang diam*bil* dari distribusi uniform $U(0,1)$,
x_{baru}	: sampel sintetis baru hasil interpolasi x_i dan x_j

Proses over*sampling* ini hanya dilakukan pada data latih agar tidak terjadi kebocoran informasi ke data uji. Dengan demikian, diharapkan model *SVM* dan *Random Forest* menjadi lebih sensitif terhadap kelas hepatitis dan mampu memberikan performa yang lebih seimbang antara kelas *positif* dan *negatif*. Selain itu, perbandingan antara skenario tanpa *SMOTE* (*baseline*) dan dengan *SMOTE* dilakukan secara terpisah untuk menilai secara kuantitatif dampak penyeimbangan kelas terhadap kinerja model.

Data Mining

Tahap ini dilakukan setelah data hepatitis melalui proses praproses dan pembagian menjadi data latih dan data uji. Pada tahap ini, data latih dimanfaatkan untuk membangun dua model klasifikasi utama, yaitu *Support Vector Machine* (*SVM*) dan *Random Forest*.*Support Vector Machine* digunakan untuk mencari hiperbidang pemisah dengan margin maksimum sehingga mampu membedakan kelas hepatitis dan non-hepatitis berdasarkan kombinasi atribut klinis dan laboratorik yang bersifat multidimensi. Secara mekanisme, *SVM* bekerja dengan mengidentifikasi titik-titik data yang paling dekat dengan batas antar kelas, yang dikenal sebagai support vector. Model kemudian membentuk hiperbidang yang memaksimalkan jarak antara kedua kelas tersebut sehingga batas keputusan yang dihasilkan lebih sta*bil* dan tidak mudah terpengaruh oleh variasi kecil pada data. Pada dataset hepatitis yang memiliki banyak parameter biokimia hati dengan pola hubungan yang kompleks, pendekatan ini memungkinkan model membangun pemisahan kelas yang optimal, baik pada pola linier maupun nonlinier melalui penggunaan fungsi kernel, fungsi keputusan *SVM* untuk kasus klasifikasi biner dapat dituliskan sebagai:

$$f(x) = \text{sign}(w^\top \varphi(x) + b) \quad (6)$$

Dimana:

X	: vektor fitur satu pasien (berisi atribut demografis dan laboratorik).
φ(x)	: transformasi fitur ke ruang berdimensi lebih tinggi (implisit melalui kernel).
w	: vektor bobot yang dipelajari oleh *SVM*.
b	: bias (intersep) dari hiperbidang pemisah.
$f(x)$	: skor keputusan sebelum diam*bil* kelas.
sign(·)	: fungsi tanda yang menentukan kelas akhir (hepatitis / non-hepatitis).

Sementara itu, *Random Forest* diterapkan sebagai model berbasis sejumlah pohon keputusan yang dibangun dari berbagai subset data dan subset fitur, kemudian menghasilkan prediksi akhir melalui agregasi keputusan seluruh pohon. Secara mekanisme, *Random Forest* membangun sejumlah pohon keputusan dari subset data latih yang dipilih secara acak melalui *bootstrap sampling*, dengan pemilihan fitur acak pada setiap pohon. Keragaman struktur pohon ini membuat model tidak bergantung pada satu pola tertentu. Prediksi akhir ditentukan melalui *voting* mayoritas, sehingga klasifikasi hepatitis dan non-hepatitis menjadi lebih sta*bil*, tahan terhadap noise, serta mampu menangkap hubungan nonlinier antar parameter biokimia hatis, prediksi akhir *Random Forest* dapat dinyatakan sebagai *voting* mayoritas dari himpunan pohon keputusan berikut:

$$\hat{y}(x) = mode\{h_b(x)\}_{b=1}^{B} \quad (7)$$

Dimana:

$\hat{y}(x)$	: prediksi akhir model *Random Forest* untuk sampel x.
x	: vektor fitur satu data pasien.
$h_b(x)$	: prediksi pohon keputusan ke-b dalam *Random Forest*.
B	: jumlah total pohon dalam *Random Forest*.
$mode$ (·)	: operator yang memilih kelas yang paling sering diprediksi (*voting* mayoritas).

Pendekatan ini membuat *Random Forest* relatif lebih sta*bil* terhadap variasi data serta mampu menangkap hubungan nonlinier antarf fitur tanpa asumsi distribusi yang kaku. Kedua algoritma tersebut diterapkan pada dataset yang sama dengan skema pra-pemrosesan yang identik, sehingga perbandingan kinerja antara *Support Vector Machine* dan *Random Forest* dalam konteks klasifikasi hepatitis dapat dilakukan secara objektif dan terukur.

Evaluasi

Evaluasi kinerja model dilakukan menggunakan *skema 10-fold cross validation* sebagaimana digambarkan pada alur penelitian di Gambar 1. Dataset dibagi menjadi sepuluh lipatan (*fold*) yang proporsional terhadap distribusi kelas, kemudian pada setiap iterasi sembilan *fold* digunakan sebagai data latih dan satu *fold* sebagai data uji, hingga seluruh *fold* bergantian berperan sebagai data uji. Prosedur ini diterapkan baik pada skenario *baseline* (tanpa *SMOTE*) maupun pada skenario setelah penerapan *SMOTE*, untuk masing-masing algoritma *Support Vector Machine* dan *Random Forest*. Pada setiap *fold*, nilai metrik evaluasi (misalnya akurasi, presisi, *recall*, dan *F1-score*) dihitung secara terpisah sehingga diperoleh deret nilai $M_1, M_2, \dots, M_{10}$ untuk setiap metrik M. Nilai rata-rata metrik $\bar{M}$digunakan sebagai indikator utama kinerja model dan didefinisikan sebagai:

$$\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M_k\,, \quad K = 10 \tag{8}$$

Dimana:

M_k : nilai metrik pada *fold* ke-k.
K : jumlah *fold* (dalam penelitian ini $K = 10$)
$\bar{M}$: nilai rata-rata metrik di seluruh *fold*.

Dengan demikian, nilai metrik dari kesepuluh lipatan kemudian dirata-ratakan untuk memperoleh gambaran performa yang lebih sta*bil* dan tidak bergantung pada satu pembagian data tertentu. Hasil evaluasi ini menjadi dasar untuk membandingkan efektivitas *SVM* dan *Random Forest*, baik sebelum maupun sesudah penerapan *SMOTE*, dalam klasifikasi hepatitis.

HASIL DAN PEMBAHASAN

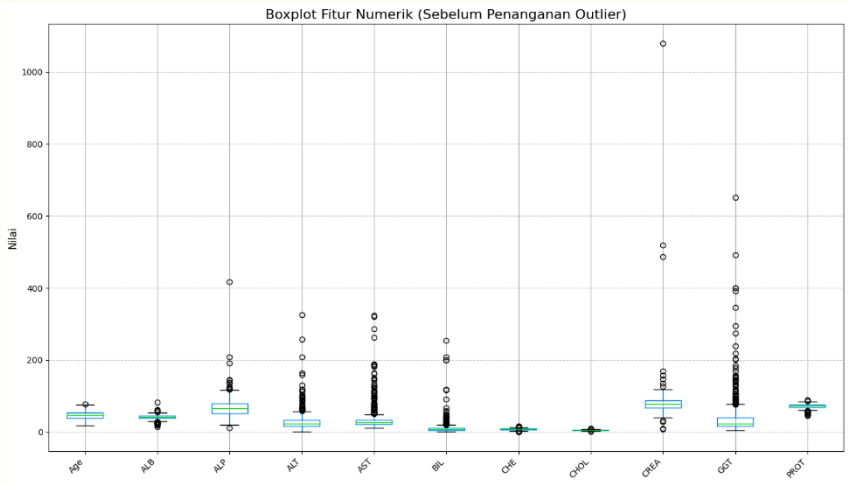
Preprocessing Data

Pada tahap preprocessing, dilakukan pengecekan kualitas data pada *dataset* hepatitis yang meliputi pemeriksaan *missing value*, identifikasi *outlier*, dan transformasi data. Dari total 615 record, tidak ditemukan data duplikat, namun beberapa parameter laboratorik seperti ALB, ALP, ALT, CHOL, dan PROT memiliki total 31 nilai hilang sehingga dilakukan imputasi menggunakan *median* untuk menjaga kualitas data tanpa mengurangi jumlah sampel. Selanjutnya, *outlier* pada fitur numerik diidentifikasi menggunakan *interquartile range* (IQR) dan ditangani dengan teknik *capping* agar distribusi data lebih stabil. Transformasi data juga dilakukan dengan memastikan label *Category* berada dalam bentuk biner (0 untuk non-hepatitis dan 1 untuk hepatitis), serta mengubah atribut *Sex* menjadi numerik 0 dan 1. Perbandingan jumlah *missing value* sebelum dan sesudah imputasi ditunjukkan pada Tabel 2.

Tabel 2. Perbandingan Missing Value Sebelum dan Sesudah Imputasi

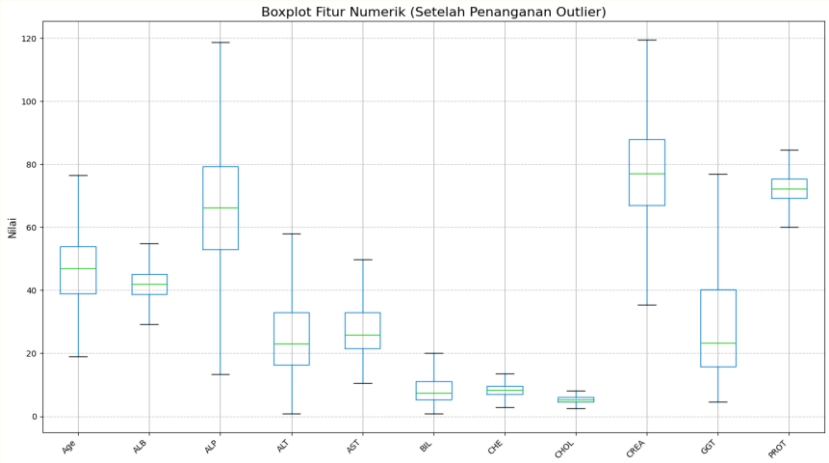
Atribut	Sebelum Penanganan	Sesudah Penanganan
Category	0	0
Age	0	0
Sex	0	0
ALB	1	0
ALP	18	0
ALT	1	0
AST	0	0
BIL	0	0
CHE	0	0
CHOL	10	0
CREA	0	0
GGT	0	0
PROT	1	0

Berdasarkan Tabel 2, *missing value* hanya ditemukan pada atribut ALB, ALP, ALT, CHOL, dan PROT dengan total 31 data hilang. Setelah dilakukan imputasi menggunakan *median*, seluruh atribut tidak lagi memiliki nilai hilang sehingga *dataset* siap digunakan pada tahap analisis berikutnya, yaitu identifikasi *outlier*. Pada tahap ini, dilakukan visualisasi awal terhadap seluruh fitur numerik menggunakan boxplot sebagaimana ditunjukkan pada Gambar 2. Hampir semua parameter biokimia hati, seperti Age, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, dan PROT, menunjukkan sebaran nilai yang cukup lebar dengan sejumlah titik data berada di luar rentang interkuartil. Kondisi tersebut mengindikasikan adanya nilai ekstrem pada beberapa fitur laboratorik yang berpotensi memengaruhi proses klasifikasi, sehingga dilakukan penanganan *outlier* menggunakan pendekatan *interquartile range* (IQR) dan teknik *capping* agar distribusi data menjadi lebih stabil.



Gambar 2. Boxplot Sebelum Penanganan *Outlier*

Setelah prosedur penanganan *outlier* diterapkan dengan pendekatan *interquartile range* (*IQR*) dan teknik *capping*, pola sebaran fitur-fitur numerik tampak jauh lebih terkonsolidasi sebagaimana ditunjukkan pada Gambar 3. Boxplot untuk parameter *Age*, *ALB*, *ALP*, *ALT*, AST, *BIL*, *CHE*, *CHOL*, *CREA*, *GGT*, dan *PROT* menunjukkan bahwa rentang antar kuartil menjadi lebih proporsional dan whisker tidak lagi didominasi oleh titik-titik ekstrem seperti pada kondisi sebelum penanganan *outlier*. Nilai-nilai yang sebelumnya berada jauh di luar batas bawah maupun batas atas telah disesuaikan ke dalam rentang yang masih wajar secara statistik, sehingga bentuk distribusi tiap fitur menjadi lebih simetris dan bebas dari pengaruh nilai ekstrim yang berlebihan. Kondisi ini mengindikasikan bahwa data numerik hasil praproses lebih sta*bil* dan representatif terhadap pola mayoritas pasien, sehingga diharapkan dapat mendukung pembentukan model klasifikasi hepatitis yang lebih robust pada tahap pemodelan menggunakan *Support Vector Machine* dan *Random Forest*.



Gambar 3. Boxplot Sesudah Penanganan *Outlier*

Setelah tahap penanganan outlier selesai, dilakukan transformasi data pada atribut Category dan Sex agar seluruh variabel berada dalam format numerik yang sesuai untuk proses pemodelan. Perubahannilai dapat dilihat pada Tabel 3.

Tabel 3. Transformasi Data

Atribut	**Nilai Kategori**	
	Sebelum	**Sesudah**
Category	0=Blood Donor	**0**
	0s=suspect Blood Donor	
	1=Hepatitis	**1**
	2=Fibrosis	
	3=Cirrhosis	
Sex	M	**0**
	F	**1**

Pada tahap transformasi label, variabel *Category* yang semula memiliki lima kategori, yaitu *0 = Blood Donor*, *0s = suspect Blood Donor*, *1 = Hepatitis*, *2 = Fibrosis*, dan *3 = Cirrhosis*, disederhanakan menjadi label biner sebagaimana dirangkum pada Tabel 3. Untuk menyelaraskan dengan tujuan penelitian yang berfokus pada pembedaan antara subjek sehat dan subjek yang mengalami gangguan hati, seluruh kategori yang merepresentasikan kondisi penyakit (Hepatitis, Fibrosis, dan Cirrhosis) dikelompokkan ke dalam kelas 1 (hepatitis), sedangkan kategori Blood Donor dan *suspect Blood Donor* dipetakan ke kelas 0 sebagai kelompok non-hepatitis. Selain itu, atribut kategorikal *Sex* yang semula direpresentasikan dengan kode huruf *M* dan *F* juga dikonversi ke bentuk numerik, masing-masing menjadi 0 untuk laki-laki dan 1 untuk perempuan. Transformasi ini menghasilkan variabel target dan fitur kategorikal yang sudah sepenuhnya numerik, sehingga lebih mudah diolah oleh algoritma *Support Vector Machine* dan *Random Forest* pada tahap pemodelan.

Pemodelan

Setelah seluruh tahapan *preprocessing* data yang meliputi penanganan *missing value*, analisis *outlier*, dan transformasi data selesai dilakukan, langkah berikutnya adalah mengevaluasi kinerja model klasifikasi yang dibangun. Pada tahap ini, algoritma *Support Vector Machine* dan *Random Forest* yang telah dilatih dengan *skema 10-fold cross validation* dinilai berdasarkan empat metrik utama, yaitu akurasi, presisi, *recall*, dan *F1-score*, baik pada skenario *baseline* (tanpa *SMOTE*) maupun pada skenario setelah penerapan *SMOTE* pada data latih. Nilai masing-masing metrik dihitung dari hasil prediksi model pada setiap *fold*, kemudian dirata-ratakan untuk memperoleh gambaran performa yang lebih sta*bil* dan tidak bergantung pada satu pembagian data tertentu. Ringkasan hasil pemodelan kedua algoritma pada skenario baseline ditampilkan pada Tabel 4 untuk SVM dan Tabel 5 untuk Random Forest.

Tabel 4. *SVM* Baseline

Fold	**Akurasi**	**Presisi**	***Recall***	**F1_Score**
1	1.0000	1.0000	1.0000	1.0000
2	0.9400	0.9600	0.9400	0.9449
3	0.9796	0.9801	0.9796	0.9788
4	0.9592	0.9694	0.9592	0.9616
5	0.9796	0.9825	0.9796	0.9803
6	0.9796	0.9825	0.9796	0.9803
7	1.0000	1.0000	1.0000	1.0000
8	0.9388	0.9428	0.9388	0.9296
9	0.9592	0.9610	0.9592	0.9556
10	0.9592	0.9610	0.9592	0.9556
Rata-Rata	0.9695	0.9739	0.9695	0.9687

Penerapan algoritma *Support Vector Machine* pada skenario *baseline* dengan skema *10-fold cross-validation* memberikan kinerja rata-rata dengan akurasi sebesar 96,95%, presisi 97,39%, *recall* 96,95%, dan *F1-score* 96,87%, sebagaimana tersaji pada Tabel 4.

Tabel 5. *Random Forest* Baseline

Fold	Akurasi	Presisi	*Recall*	F1_Score
1	1.0000	1.0000	1.0000	1.0000
2	0.9400	0.9600	0.9400	0.9449
3	0.9592	0.9610	0.9592	0.9556
4	1.0000	1.0000	1.0000	1.0000
5	0.9592	0.9610	0.9592	0.9556
6	0.9184	0.9109	0.9184	0.9111
7	1.0000	1.0000	1.0000	1.0000
8	0.8980	0.8829	0.8980	0.8827
9	0.9592	0.9610	0.9592	0.9556
10	0.9592	0.9610	0.9592	0.9556
Rata-Rata	0.9593	0.9598	0.9593	0.9561

Secara keseluruhan, penerapan algoritma *Random Forest* pada skenario *baseline* dengan skema *10-fold cross-validation* menghasilkan kinerja rata-rata dengan akurasi sebesar 95,93%, presisi 95,98%, *recall* 95,93%, dan *F1-score* 95,61%, sebagaimana tersaji pada Tabel 5. Dibandingkan dengan hasil rata-rata *SVM baseline*, seluruh metrik *Random Forest* tersebut sedikit lebih rendah, sehingga pada konfigurasi tanpa *SMOTE* ini *SVM* masih menunjukkan performa klasifikasi hepatitis yang lebih unggul.

SMOTE

Pada skenario pemodelan kedua, data latih terlebih dahulu diseimbangkan menggunakan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) sebelum dilatih dengan algoritma klasifikasi. Setelah proses over*sampling* selesai, model *Support Vector Machine* dan *Random Forest* dibangun dan dievaluasi menggunakan *skema 10-fold cross validation* dengan prosedur yang sama seperti pada skenario *baseline*. Untuk setiap kombinasi model dan *fold*, dihitung nilai akurasi, presisi, *recall*, dan *F1-score*, sehingga diperoleh rangkaian hasil kinerja per *fold* beserta nilai rata-rata di seluruh 10 *fold*. Ringkasan hasil pengujian *SVM* berbasis *SMOTE* ditampilkan pada Tabel 6, sedangkan hasil pengujian *Random Forest* dengan *SMOTE* disajikan pada Tabel 7.

Tabel 6. Hasil *SVM* Menggunakan *SMOTE*

Fold	Akurasi	Presisi	*Recall*	F1_Score
1	0.9770	0.9780	0.9770	0.9770
2	0.9540	0.9579	0.9540	0.9540
3	1.0000	1.0000	1.0000	1.0000
4	0.9540	0.9540	0.9540	0.9540
5	0.9884	0.9886	0.9884	0.9884
6	0.9884	0.9886	0.9884	0.9884
7	1.0000	1.0000	1.0000	1.0000
8	1.0000	1.0000	1.0000	1.0000
9	0.9884	0.9886	0.9884	0.9884
10	1.0000	1.0000	1.0000	1.0000
Rata-Rata	0.9850	0.9856	0.9850	0.9850

Penerapan algoritma *Support Vector Machine* pada skenario dengan *SMOTE* dan skema *10-fold cross-validation* menghasilkan kinerja rata-rata dengan akurasi sebesar 98,50%, presisi 98,56%, *recall* 98,50%, dan *F1-score* 98,50%, sebagaimana ditunjukkan pada Tabel 6.

Tabel 7. Hasil *Random Forest* Menggunakan *SMOTE*

Fold	Akurasi	Presisi	*Recall*	F1_Score
1	0.9770	0.9780	0.9770	0.9770
2	0.9770	0.9780	0.9770	0.9770
3	1.0000	1.0000	1.0000	1.0000
4	0.9655	0.9658	0.9655	0.9655
5	0.9884	0.9886	0.9884	0.9884
6	0.9884	0.9886	0.9884	0.9884
7	1.0000	1.0000	1.0000	1.0000
8	0.9884	0.9886	0.9884	0.9884
9	1.0000	1.0000	1.0000	1.0000
10	1.0000	1.0000	1.0000	1.0000
Rata-Rata	0.9885	0.9888	0.9885	0.9885

Secara keseluruhan, penerapan algoritma *Random Forest* pada skenario dengan *SMOTE* dan skema *10-fold cross-validation* menghasilkan kinerja rata-rata dengan akurasi sebesar 98,85%, presisi 98,88%, *recall* 98,85%, dan *F1-score* 98,85%, sebagaimana tersaji pada Tabel 7. Jika dibandingkan dengan hasil rata-rata *SVM* pada skenario yang sama, seluruh metrik *Random Forest* tersebut sedikit lebih tinggi, sehingga pada konfigurasi dengan penyeimbangan kelas berbasis *SMOTE* ini *Random Forest* menunjukkan performa klasifikasi hepatitis yang lebih unggul.

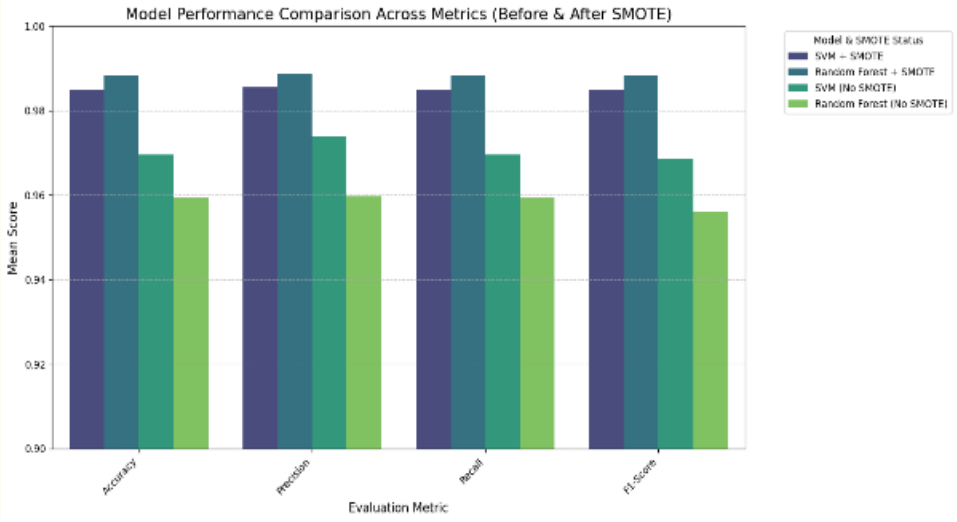
Evaluasi

Evaluasi perhitungan antar model dalam pengujian menggunakan algoritma *SVM* dan *Random Forest* dengan penambahan metode *SMOTE* dapat dibandingkan dengan nilai akurasi, presisi, *recall*, dan *F1-score*. Nilai perbandingan pengukuran tersebut ditunjukkan pada Tabel 8.

Tabel 8. Hasil Evaluasi Model

Model	Akurasi	Presisi	*Recall*	*F1-Score*
SVM + Smote	0.985017	0.985588	0.985017	0.985008
Random Forest + Smote	0.988466	0.988773	0.988466	0.988463
SVM	0.969510	0.973922	0.969510	0.968658
Random Forest	0.959306	0.959775	0.959306	0.956101

Hasil pengujian dari dua skenario eksperimen memperlihatkan perbedaan kinerja yang jelas antara pemodelan hepatitis tanpa penyeimbangan kelas (*baseline*) dan pemodelan dengan penerapan *SMOTE* pada algoritma *Support Vector Machine* dan *Random Forest*. Penerapan *SMOTE* pada data latih terbukti mampu meningkatkan nilai akurasi, presisi, *recall*, dan *F1-score* terutama pada kelas hepatitis yang sebelumnya berada dalam posisi minoritas, sekaligus memperbaiki kesta*bil*an kinerja model di seluruh 10 lipatan validasi. Secara umum, skenario dengan *SMOTE* menghasilkan metrik evaluasi yang lebih tinggi dibandingkan skenario *baseline* pada kedua algoritma, dengan peningkatan yang lebih menonjol terlihat pada *Random Forest*. Grafik perbandingan metrik evaluasi untuk kedua skenario tersebut divisualisasikan pada Gambar 4.



Gambar 4. Perbandingan Kinerja Model Baseline dan SMOTE

KESIMPULAN

Berdasarkan rangkaian tahapan penelitian mulai dari preprocessing data, penerapan SMOTE, hingga pembangunan model klasifikasi menggunakan Support Vector Machine dan Random Forest dengan skema 10-fold cross validation, dapat disimpulkan bahwa pipeline yang digunakan mampu menghasilkan performa klasifikasi hepatitis yang tinggi dan stabil. Pada skenario baseline, SVM menunjukkan kinerja sedikit lebih baik dibandingkan Random Forest dengan rata-rata akurasi di atas 96%. Setelah penerapan SMOTE, kedua algoritma mengalami peningkatan performa, dengan Random Forest memberikan hasil terbaik melalui akurasi 98,85%, presisi 98,88%, recall 98,85%, dan F1-score 98,85%. Hasil ini menunjukkan bahwa penyeimbangan kelas menggunakan SMOTE efektif meningkatkan sensitivitas model terhadap kelas hepatitis, sementara kontribusi penelitian terletak pada penggunaan pipeline preprocessing dan evaluasi yang seragam untuk membandingkan dampak SMOTE secara langsung pada dua algoritma klasifikasi hepatitis. Penelitian selanjutnya dapat diarahkan pada optimasi hyperparameter, penambahan algoritma lain, serta validasi pada dataset klinis yang lebih beragam.

DAFTAR PUSTAKA

Ali, A. M., Hassan, M. R., Aburub, F., Alauthman, M., Aldweesh, A., Al-Qerem, A., Jebreen, I., & Nabot, A. (2023). Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection. *Machines*, *11*(3), 1–14. https://doi.org/10.3390/machines11030391

Alnur, B., Mulyono, M., Amillia, F., & Sutoyo, S. (2023). JITE (Journal of Informatics and Telecommunication Engineering). *Journal of Informatics and Telecommunication Engineering*, *7*(1), 102–111. https://doi.org/10.31289/jite.v8i2.13218 Received:

Aurelia, J. E., Rustam, Z., Wirasati, I., Hartini, S., & Saragih, G. S. (2021). Hepatitis classification using support vector machines and random forest. *IAES International Journal of Artificial Intelligence*, *10*(2), 446–451. https://doi.org/10.11591/IJAI.V10.I2.PP446-451

Barata, B., Noersasongko, M. E., Purwanto, M. A. S. (2021). Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using Electronic Nose. *Resti*, *1*(10), 19–25. https://doi.org/10.29207/resti.v7i2.4687

Barata, M., Dwi Irnawati, Ifnu Wisma Dwi Prastya, & Dwi Issadari Hastuti. (2025). Hydrogen Sulfide Leak Detection Using the C4.5 Algorithm: Optimizing Feature Extraction for Enhanced Accuracy.

PROCEEDING AL GHAZALI International Conference, *2*, 348–358. https://doi.org/10.52802/aicp.v1i1.1352

Cabanillas-Carbonell, M., & Zapata-Paulini, J. (2025). Seeking best performance: a comparative evaluation of machine learning models in the prediction of hepatitis C. *Indonesian Journal of Electrical Engineering and Computer Science*, *39*(1), 374. https://doi.org/10.11591/ijeecs.v39.i1.pp374-386

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 677–690. https://doi.org/10.30812/matrik.v21i3.1726

Gunawan, R. G., & Ilham Pratama, M. (2024). Analisa Kinerja Algoritma Machine Learning Untuk Prediksi Virus Hepatitis C. *Jurnal CoSciTech (Computer Science and Information Technology)*, *4*(3), 772–777. https://doi.org/10.37859/coscitech.v4i3.6513

Kementerian Kesehatan RI. (2018). Riset Kesehatan Dasar (Riskedas). Laporan Nasional Riskesdad.2018.Kementerian Kesehatan RI. Badan Penelitian dan Pengembangan Kesehatan. *Laporan Nasional Riskesndas 2018*, *44*(8), 181–222. http://www.yankes.kemkes.go.id/assets/downloads/PMK No. 57 Tahun 2013 tentang PTRM.pdf

Kementrian Kesehatan. (2016). *Profil Kesehatan*.

Khairunnas, Masitha, A., & Rafiuddin. (2025). Identifikasi Klaster UMKM di Kota Bima menuju Indonesia Emas 2045 dengan Metode Support Vector Machine. *Decode: Jurnal Pendidikan Teknologi Informasi*, *5*(3), 967–981. https://doi.org/10.51454/decode.v5i3.1359

Khatun, P., Umam, S., Razzak, R. B., Shamsuddin, I. B., & Salma, N. (2025). A study on the effectiveness of machine learning models for hepatitis prediction. *Scientific Reports*, *15*(1), 1–16. https://doi.org/10.1038/s41598-025-07104-4

Kumari, S., Das, S., Sonker, P. K., Saroj, A., & Kumar, M. (2025). Prediction of hepatitis-C virus using statistical learning models. *Discover Public Health*, *22*(1). https://doi.org/10.1186/s12982-025-00654-y

Lilhore, U. K., Manoharan, P., Sandhu, J. K., Simaiya, S., Dalal, S., Baqasah, A. M., Alsafyani, M., Alroobaea, R., Keshta, I., & Raahemifar, K. (2023). Hybrid model for precise hepatitis-C classification using improved random forest and SVM method. *Scientific Reports*, *13*(1), 1–18. https://doi.org/10.1038/s41598-023-36605-3

Farghaly, H. M., Shams, M. Y., & El-Hafeez, T. A (2023). Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. *Knowledge and Information Systems*, *65*(6), 2595–2617. https://doi.org/10.1007/s10115-023-01851-4

Mehzabeen, S. M., Gayathri, R., Paramasaivam, P., & Ramya, A. (2025). Enhancing Hepatitis C Diagnosis: The Impact of SMOTE, Optuna, and SHAP on Detection Methods. *Iranian Journal of Electrical and Electronic Engineering*, *21*(4), 1–16. https://doi.org/10.22068/IJEEE.21.4.3418

Nugraha, M. A., Mazdadi, M. I., Farmadi, A., Muliadi, & Saragih, T. H. (2023). Penyeimbangan Kelas SMOTE dan Seleksi Fitur Ensemble Filter pada Support Vector Machine untuk Klasifikasi Penyakit Liver. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(6), 1273–1284. https://doi.org/10.25126/jtiik.2023107234

Purnomo, A., Barata, M. A., Soeleman, M. A., & Alzami, F. (2020). Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease. *Journal of Physics: Conference Series*, *1511*(1), 1–7. https://doi.org/10.1088/1742-6596/1511/1/012001

Putri, S. A., & Rachmatika, R. (2025). Penerapan Algoritma Random Forest dan SMOTE untuk Prediksi Risiko Putus Sekolah Siswa Sekolah Menengah Kejuruan. *DECODE: Jurnal Pendidikan Teknologi*

Informasi, *5*(3), 903–910. https://doi.org/Doi: http://dx.doi.org/10.51454/decode.v5i3.1360

Rehman, A. U., Butt, W. H., Ali, T. M., Javaid, S., Almufareh, M. F., Humayun, M., Rahman, H., Mir, A., & Shaheen, M. (2024). A Machine Learning-Based Framework for Accurate and Early Diagnosis of Liver Diseases: A Comprehensive Study on Feature Selection, Data Imbalance, and Algorithmic Performance. *International Journal of Intelligent Systems*, *2024*, 1–29. https://doi.org/10.1155/2024/6111312

Sharfina, N., & Ramadhan, N. G. (2023). Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes. *JOINTECS (Journal of Information Technology and Computer Science)*, *8*(1), 33. https://doi.org/10.31328/jointecs.v8i1.4456

Syukron, M., Santoso, R., & Widiharih, T. (2020). PADA IMBALANCE CLASS DATA Muhamad. *JURNAL GAUSSIAN*, *9*, 227–236.

Yaqin, A. A., Barata, M. A., & Mahmudah, N. (2025). Implementation of the Random Forest Algorithm with Optuna Optimization in Lung Cancer Classification. *Sistemasi*, *14*(2), 561. https://doi.org/10.32520/stmsi.v14i2.4877

Zhang, S., & Cui, F. (2025). Global progress, challenges and strategies in eliminating public threat of viral hepatitis. *Infectious Diseases of Poverty*, *14*(1), 25–28. https://doi.org/10.1186/s40249-025-01275-y