Comparative Classification of Hepatitis C Disease Using Naïve Bayes and Random Forest with SMOTE-Based Class Balancing

Ida Oktavia Salsavana[1*], Mokhamad Amin Hariyadi[2], Fresy Nugroho[3]
[1-3]Program Studi Magister Informatika, Universitas Islam Negri Maulana Malik Ibrahim Malang
Corresponding Author's e-mail : *Idasalsa234@gmail.com*

ARMADA
JURNAL PENELITIAN MULTIDISIPLIN

e-ISSN: 2964-2981
ARMADA : Jurnal Penelitian Multidisiplin
https://ejournal.45mataram.ac.id/index.php/armada
Vol. 04, No. 06 Juni, 2026
Page: 2319-2331

DOI:
https://doi.org/10.55681/armada.v4i6.2961

Article History:
Received: April 08, 2026
Revised: Mei 13, 2026
Accepted: Juni 18, 2026

Abstract *: Hepatitis C is a liver disease caused by the Hepatitis C Virus (HCV) that can lead to serious complications such as cirrhosis, liver failure, and liver cancer. Early and accurate detection is crucial to improving treatment outcomes and patient quality of life. This study compares the performance of Naïve Bayes and Random Forest algorithms in classifying Hepatitis C disease using clinical data from the UCI Machine Learning Repository, consisting of 615 patient records. The preprocessing stage included data cleaning, data transformation, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. The dataset was divided into training and testing sets at an 80:20 ratio. Results show that Random Forest achieved an accuracy of 99.06%, precision of 0.99, recall of 0.98, and F1-score of 0.99, outperforming Naïve Bayes which obtained an accuracy of 84.06%. Feature importance analysis identified AST, ALT, GGT, Age, and Albumin as the most significant clinical predictors. The combination of Random Forest and SMOTE proves to be a highly effective approach for Hepatitis C classification and holds strong potential to support accurate clinical decision-making in early disease detection.*

Keywords : *Class Imbalance, Hepatitis C, Naïve Bayes, Random Forest, SMOTE*

Abstrak : Hepatitis C merupakan penyakit hati yang disebabkan oleh Virus Hepatitis C (HCV) dan dapat berkembang menjadi komplikasi serius seperti sirosis, gagal hati, dan kanker hati. Deteksi dini yang akurat sangat penting untuk meningkatkan keberhasilan pengobatan dan kualitas hidup pasien. Penelitian ini membandingkan kinerja algoritma Naïve Bayes dan Random Forest dalam klasifikasi penyakit Hepatitis C menggunakan data klinis dari UCI Machine Learning Repository yang terdiri atas 615 data pasien. Tahap prapemrosesan meliputi pembersihan data, transformasi data, dan penyeimbangan kelas menggunakan metode Synthetic Minority Over-sampling Technique (SMOTE). Dataset dibagi dengan rasio 80:20 untuk data latih dan data uji. Hasil eksperimen menunjukkan bahwa Random Forest menghasilkan akurasi 99,06%, precision 0,99, recall 0,98, dan F1-score 0,99, jauh melampaui Naïve Bayes yang memperoleh akurasi 84,06%. Analisis feature importance mengidentifikasi AST, ALT, GGT, Usia, dan Albumin sebagai prediktor klinis paling signifikan. Kombinasi algoritma Random Forest dan SMOTE terbukti menjadi pendekatan yang sangat efektif dalam klasifikasi Hepatitis C dan berpotensi besar mendukung pengambilan keputusan klinis yang akurat untuk deteksi

dini penyakit.

Kata Kunci : Hepatitis C, Ketidakseimbangan Kelas, Naïve Bayes, Random Forest, SMOTE

PENDAHULUAN

Hepatitis C merupakan penyakit hati yang disebabkan oleh infeksi Virus Hepatitis C (HCV) dan masih menjadi salah satu masalah kesehatan masyarakat di dunia (World Health Organization, 2024; Blach et al., 2023). Jutaan orang hidup dengan infeksi Hepatitis C kronis, dan sebagian besar kasus tidak terdeteksi pada tahap awal karena gejalanya sering kali tidak muncul secara signifikan (World Health Organization, 2024). Apabila tidak ditangani dengan baik, Hepatitis C kronis dapat berkembang menjadi sirosis hati, gagal hati, bahkan kanker hati yang dapat meningkatkan angka kesakitan dan kematian (Blach et al., 2023; Lingala & Ghany, 2022). Oleh karena itu, deteksi dini dan diagnosis yang akurat menjadi faktor penting dalam meningkatkan keberhasilan pengobatan dan kualitas hidup pasien (Lingala & Ghany, 2022).

Perkembangan teknologi informasi, khususnya di bidang *data mining* dan *machine learning*, telah membuka peluang baru dalam mendukung proses diagnosis penyakit (Han et al., 2022; Shameer et al., 2022). Algoritma *machine learning* mampu menganalisis data klinis dalam jumlah besar, menemukan pola tersembunyi, serta membantu tenaga medis dalam pengambilan keputusan (Shameer et al., 2022; Ahmed et al., 2023). Dalam bidang kesehatan, teknik klasifikasi telah banyak digunakan untuk memprediksi berbagai penyakit, seperti diabetes, penyakit jantung, penyakit hati, dan Hepatitis C (Ahmed et al., 2023; Arslan et al., 2022).

Di antara berbagai algoritma klasifikasi, *Naïve Bayes* dan *Random Forest* merupakan metode yang banyak digunakan karena memiliki karakteristik yang berbeda namun sama-sama efektif dalam proses klasifikasi (Zhang, 2021; Breiman, 2001). *Naïve Bayes* merupakan algoritma berbasis probabilitas yang sederhana, cepat, dan efisien dalam mengolah data berdimensi tinggi (Zhang, 2021). Sementara itu, *Random Forest* merupakan metode *ensemble learning* yang mampu menangani hubungan kompleks antarvariabel dan memiliki kemampuan generalisasi yang baik (Breiman, 2001; Probst et al., 2019). Beberapa penelitian sebelumnya menunjukkan bahwa kedua algoritma tersebut mampu memberikan performa yang baik dalam klasifikasi penyakit Hepatitis C (Arslan et al., 2022; Furizal et al., 2023).

Meskipun demikian, salah satu tantangan utama dalam klasifikasi Hepatitis C adalah ketidakseimbangan kelas (*class imbalance*) (Haixiang et al., 2017). Pada dataset medis, jumlah data pasien sehat sering kali jauh lebih banyak dibandingkan jumlah data pasien yang terdiagnosis Hepatitis C atau penyakit hati lainnya. Kondisi ini dapat menyebabkan model klasifikasi cenderung memprediksi kelas mayoritas sehingga menurunkan kemampuan model dalam mengenali kelas minoritas (Haixiang et al., 2017; Chawla et al., 2002). Untuk mengatasi permasalahan tersebut, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) banyak digunakan untuk menyeimbangkan distribusi data dengan menghasilkan sampel sintetis pada kelas minoritas (Chawla et al., 2002; Fernández et al., 2018).

Berbagai penelitian telah menerapkan algoritma *machine learning* untuk prediksi dan klasifikasi Hepatitis C (Arslan et al., 2022; Furizal et al., 2022; Ghosh et al., 2021). Namun, hasil yang diperoleh masih bervariasi tergantung pada karakteristik dataset, metode klasifikasi yang digunakan, serta teknik penanganan ketidakseimbangan data yang diterapkan. Selain itu, masih terbatas penelitian yang secara khusus membandingkan kinerja algoritma *Naïve Bayes* dan *Random Forest* setelah penerapan SMOTE pada dataset Hepatitis C. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengevaluasi efektivitas kedua algoritma tersebut dalam kondisi data yang telah diseimbangkan.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membandingkan kinerja algoritma *Naïve Bayes* dan *Random Forest* dalam klasifikasi penyakit Hepatitis C menggunakan dataset klinis yang tersedia secara publik. Teknik SMOTE diterapkan untuk mengatasi ketidakseimbangan kelas dan meningkatkan kemampuan model dalam mendeteksi kelas minoritas. Kontribusi penelitian ini terletak pada evaluasi pengaruh SMOTE terhadap performa klasifikasi serta identifikasi variabel klinis yang paling berpengaruh dalam klasifikasi

Hepatitis C. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan model *machine learning* yang lebih andal untuk mendukung diagnosis dini Hepatitis C.

LITERATURE REVIEW

Berbagai penelitian terdahulu telah menerapkan algoritma machine learning untuk klasifikasi dan prediksi penyakit Hepatitis C. Ringkasan penelitian terdahulu yang relevan disajikan pada Tabel 1.

Tabel 1. Ringkasan Penelitian Terdahulu

No	Peneliti	Judul Penelitian	Hasil Penelitian	Kesenjangan/Keterbatasan
1	Safdari et al. (2022)	Klasifikasi Hepatitis C menggunakan berbagai algoritma machine learning dengan SMOTE	Random Forest memberikan performa terbaik dengan akurasi 97,29% dan AUC 0,998 setelah penerapan SMOTE.	Tidak melakukan analisis khusus mengenai perbandingan kinerja Naïve Bayes dan Random Forest setelah penerapan SMOTE.
2	Lilhore et al. (2023)	Model Hibrida Improved Random Forest dan Support Vector Machine untuk Klasifikasi Hepatitis C	Kombinasi Improved Random Forest dan SVM menghasilkan performa lebih baik dibandingkan algoritma tunggal.	Tidak mengevaluasi performa algoritma Naïve Bayes sebagai metode pembanding.
3	Fan et al. (2023)	Prediksi Hepatitis C Menggunakan Random Forest dan Explainable Artificial Intelligence	Random Forest yang dioptimasi memperoleh akurasi 99,44% dan AUC 0,9986.	Berfokus pada interpretabilitas model dan tidak membahas pengaruh teknik penyeimbangan data secara mendalam.
4	I Gede Hendrayana et al. (2025)	Implementasi Random Forest dengan Teknik Resampling untuk Klasifikasi Hepatitis C	Random Forest dengan SMOTE+ENN menghasilkan akurasi 99,19% dan ROC-AUC 0,9999.	Hanya berfokus pada Random Forest dan belum membandingkannya dengan Naïve Bayes.

Berdasarkan penelitian terdahulu, algoritma Random Forest menunjukkan performa yang sangat baik dalam klasifikasi penyakit Hepatitis C. Beberapa penelitian juga membuktikan bahwa teknik penyeimbangan data seperti SMOTE mampu meningkatkan kemampuan model dalam mendeteksi kelas minoritas. Namun, sebagian besar penelitian lebih berfokus pada pengembangan model Random Forest tanpa melakukan perbandingan mendalam dengan algoritma Naïve Bayes pada data yang telah diseimbangkan. Oleh karena itu, penelitian ini dilakukan untuk membandingkan kinerja algoritma Naïve Bayes dan Random Forest pada dataset Hepatitis C yang telah melalui proses penyeimbangan kelas menggunakan SMOTE. Perbandingan dilakukan berdasarkan nilai accuracy, precision, recall, dan F1-score untuk memperoleh pemahaman yang lebih komprehensif mengenai efektivitas kedua algoritma.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan *data mining* untuk klasifikasi penyakit Hepatitis C dengan membandingkan algoritma *Naïve Bayes* dan *Random Forest*. Dataset klinis yang digunakan melalui tahap *data cleaning*, transformasi data, dan penyeimbangan kelas menggunakan SMOTE. Selanjutnya, data dibagi menjadi data latih dan data uji untuk membangun model klasifikasi. Kinerja kedua algoritma dievaluasi menggunakan *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *ROC-AUC* untuk menentukan metode yang memberikan hasil klasifikasi terbaik. Hasil evaluasi kemudian dianalisis untuk membandingkan efektivitas *Naïve Bayes* dan *Random Forest* dalam diagnosis Hepatitis C.

A. Data Preprocessing

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum proses klasifikasi. Langkah-langkah yang dilakukan meliputi:

- *Data Cleaning*, yaitu mengatasi *missing value* dan inkonsistensi data.
- *Data Transformation*, yaitu mengubah atribut kategorikal menjadi bentuk numerik.

- *Data Balancing* menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan kelas pada dataset.

B. Model Development

Dua algoritma klasifikasi digunakan dalam penelitian ini:

Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dengan asumsi independensi antar atribut. Algoritma ini menghitung probabilitas posterior setiap kelas berdasarkan data pelatihan (Zhang, 2021).

Random Forest

Random Forest merupakan metode *ensemble learning* yang membangun banyak *decision tree* menggunakan teknik *bootstrap sampling* dan *random feature selection*. Hasil klasifikasi ditentukan berdasarkan mekanisme *majority voting* dari seluruh pohon keputusan (Breiman, 2001).

C. Experimental Design

Dataset dibagi menjadi data pelatihan (*training set*) dan data pengujian (*testing set*). Kedua model dilatih menggunakan data pelatihan dan kemudian diuji menggunakan data pengujian untuk memperoleh hasil klasifikasi.

D. Performance Evaluation

Kinerja model dievaluasi menggunakan beberapa metrik, yaitu:

- *Accuracy*
- *Precision*
- *Recall*
- *F1-Score*
- *Confusion Matrix*
- *ROC Curve* dan *Area Under Curve* (AUC)

Perbandingan nilai-nilai tersebut digunakan untuk menentukan algoritma yang paling efektif dalam klasifikasi penyakit Hepatitis C.

- Dataset
- Preprocessing
- SMOTE
- *Naïve Bayes*
- *Random Forest*
- Evaluasi (*Accuracy*, *Precision*, *Recall*, *F1-Score*, *ROC-AUC*)

HASIL DAN PEMBAHASAN

A. Naïve Bayes Performance

Tahapan dalam menggunakan algoritma *Naïve Bayes* adalah sebagai berikut:

Teorema *Naïve Bayes*

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}$$

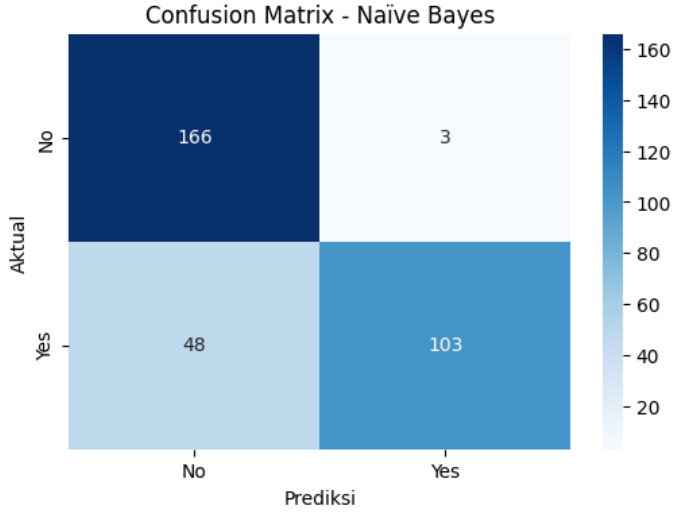
- *P(C|X)* = probabilitas kelas *C* diberikan fitur *X*
- *P(X|C)* = probabilitas mendapatkan fitur *X* diberikan kelas *C*
- *P(C)* = probabilitas dari kelas *C* (*prior probability*)
- *P(X)* = probabilitas dari fitur *X* (*evidence*)

Algoritma *Naïve Bayes* menghasilkan akurasi sebesar 84,06% dengan *precision* 0,87, *recall* 0,83, dan *F1-score* 0,83. Hasil *confusion matrix* menunjukkan bahwa dari 320 data uji, model berhasil mengklasifikasikan 165 data negatif (*True Negative*) dan 103 data positif (*True Positive*). Namun, masih terdapat 48 kasus *False Negative* yang menunjukkan bahwa beberapa pasien Hepatitis C tidak terdeteksi oleh model.

Tabel 2. Hasil Akurasi NB

	Precision	Recall	f1-score	support
0	0.78	0.98	0.87	169

1	0.97	0.68	0.80	151
Accuracy			0.84	320
macro avg	0.87	0.83	0.83	320
weighted avg	0.87	0.84	0.84	320



Gambar 1. Confusion Matrix Naïve Bayes

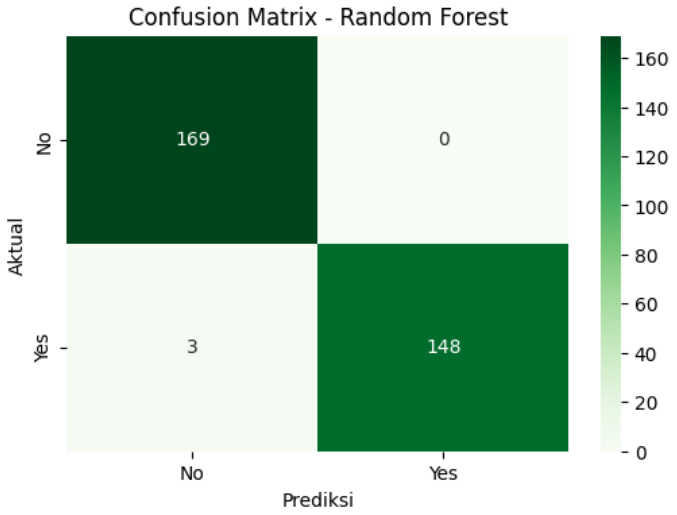
Confusion matrix menggambarkan kinerja model *Naïve Bayes* dalam mengklasifikasikan data Hepatitis C. Dari total 320 data uji, model berhasil mengklasifikasikan 166 data negatif (*True Negative*) dan 103 data positif (*True Positive*) dengan benar. Model ini terbukti sangat akurat dalam mengenali pasien sehat, ditunjukkan dengan hanya terdapat 3 kasus *False Positive* — yakni data sehat yang secara keliru diklasifikasi sebagai sakit.

B. Random Forest Performance

Random Forest menunjukkan performa yang jauh lebih baik dibandingkan *Naïve Bayes*. Model memperoleh akurasi 99,06%, *precision* 0,99, *recall* 0,98, dan *F1-score* 0,99. Dari 320 data uji, model berhasil mengklasifikasikan 169 data negatif dan 148 data positif dengan benar, serta hanya menghasilkan 3 kasus *False Negative* dan tidak terdapat *False Positive*. Hasil ini menunjukkan kemampuan *Random Forest* dalam mengenali pola kompleks pada data medis dengan tingkat kesalahan yang sangat rendah.

Tabel 3. Hasil Akurasi Random Forest

	Precision	Recall	f1-score	support
0	0.98	1.00	0.99	169
1	1.00	0.98	0.99	151
Accuracy			0.99	320
macro avg	0.99	0.99	0.99	320
weighted avg	0.99	0.99	0.99	320



Gambar 2. Confusion Matrix Random Forest

Confusion matrix dari model *Random Forest* menggambarkan performa klasifikasi terhadap data uji sebanyak 320 rekod. Model berhasil mengklasifikasikan 169 data kelas negatif (bukan Hepatitis C) dengan benar (*True Negative*) dan 148 data kelas positif (penderita Hepatitis C) dengan benar (*True Positive*). Sementara itu, terdapat 0 kasus *False Positive* dan hanya 3 kasus *False Negative*. Nilai *False Positive* yang nol menunjukkan model sangat berhati-hati sebelum mendiagnosis pasien, dan *True Positive* yang tinggi menunjukkan sensitivitas pendeteksian penyakit yang sangat baik. Secara keseluruhan, *confusion matrix* ini memperkuat bukti bahwa model *Random Forest* memberikan kinerja klasifikasi yang jauh lebih presisi dibandingkan algoritma pembandingnya.

Dataset Hepatitis C yang digunakan pada penelitian ini terdiri dari 615 data pasien. Setelah dilakukan proses *preprocessing* dan penyeimbangan kelas menggunakan metode SMOTE, jumlah data meningkat menjadi 1.066 data sehingga distribusi kelas menjadi lebih seimbang untuk proses klasifikasi. Tahapan *preprocessing* meliputi pembersihan data, transformasi atribut, dan analisis distribusi data sebelum dilakukan pelatihan model.

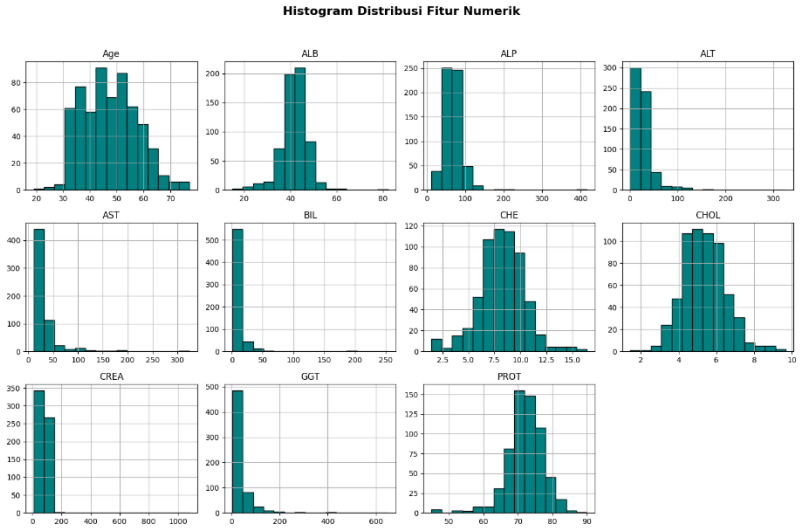
Analisis eksploratif menunjukkan bahwa beberapa atribut klinis memiliki hubungan yang kuat terhadap diagnosis Hepatitis C. Hasil visualisasi *heatmap* korelasi memperlihatkan bahwa variabel AST, ALT, GGT, *Age*, dan *Albumin* (ALB) memiliki kontribusi yang paling dominan dalam membedakan pasien sehat dan pasien Hepatitis C. Variabel-variabel tersebut merupakan indikator fungsi hati yang secara medis berhubungan erat dengan kerusakan hati akibat infeksi virus Hepatitis C.

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...
611	3=Cirrhosis	62	f	32.0	416.6	5.9	110.3	50.0	5.57	6.30	55.7	650.9	68.5
612	3=Cirrhosis	64	f	24.0	102.8	2.9	44.4	20.0	1.54	3.02	63.0	35.9	71.3
613	3=Cirrhosis	64	f	29.0	87.3	3.5	99.0	48.0	1.66	3.63	66.7	64.2	82.0
614	3=Cirrhosis	46	f	33.0	NaN	39.0	62.0	20.0	3.56	4.20	52.0	50.0	71.0
615	3=Cirrhosis	59	f	36.0	NaN	100.0	80.0	12.0	9.07	5.30	67.0	34.0	68.0

615 rows x 13 columns

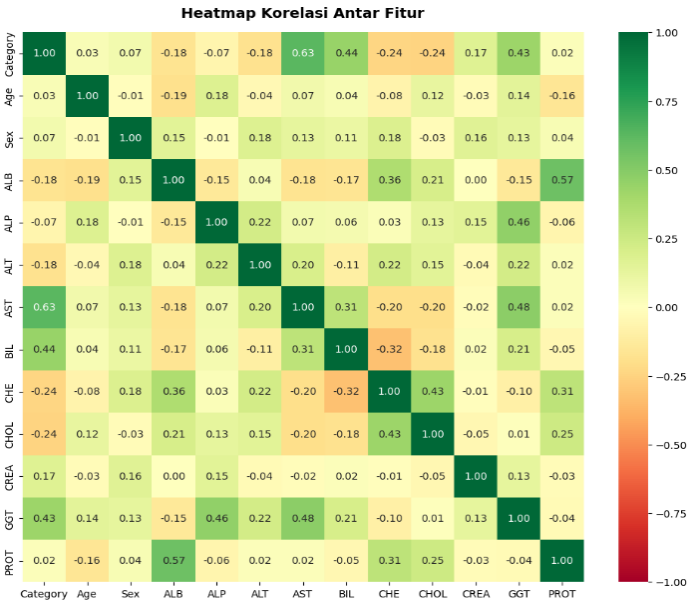
Gambar 3. Load Dataset

Dataset ini merupakan kumpulan data kesehatan yang terdiri dari 615 entri pasien (setelah melalui tahap pembersihan data) dengan 13 atribut medis yang berkaitan dengan fungsi hati individu. Kolom-kolom yang tersedia meliputi: jenis kelamin (*Sex*) yang menunjukkan apakah pasien laki-laki atau perempuan, usia (*Age*) dalam satuan tahun, serta serangkaian indikator laboratorium fungsi hati seperti kadar bilirubin (BIL), kolinesterase (CHE), *alkaline phosphatase* (ALP), protein albumin (ALB), *Aspartate Aminotransferase* (AST), *Alanine Aminotransferase* (ALT), kadar kolesterol (CHOL), kreatinin (CREA), *Gamma-Glutamyl Transferase* (GGT), dan total protein (PROT). Kolom terakhir adalah *Category*, yang berfungsi sebagai label target. Dalam penelitian ini, label target tersebut telah ditransformasi menjadi dua kelas biner: "no" (Sehat/*Blood Donor*) dan "yes" (Sakit/Hepatitis C), yang digunakan sebagai dasar untuk membangun model klasifikasi penyakit. Secara keseluruhan, dataset ini menyediakan informasi klinis yang komprehensif dan sangat relevan untuk pengembangan model klasifikasi berbasis *machine learning* dalam mendeteksi penyakit Hepatitis C secara akurat berdasarkan atribut kesehatan yang tersedia.



Gambar 4. Histogram Distribusi Hepatitis C

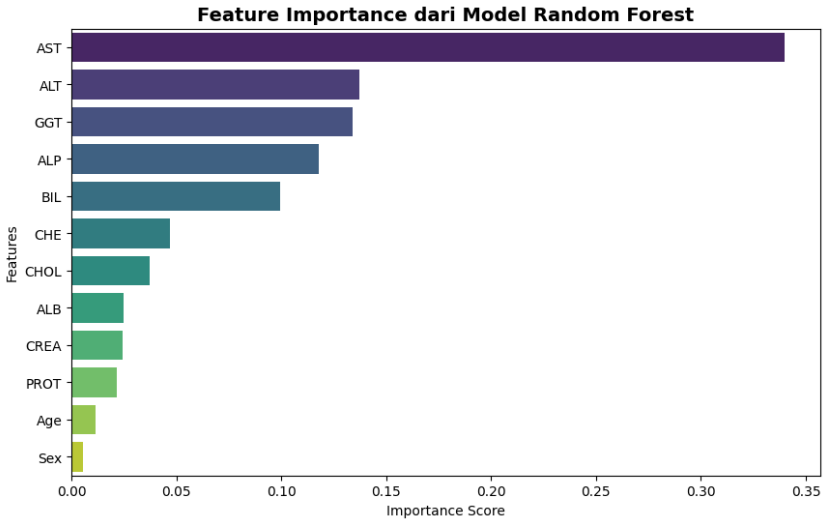
Histogram distribusi fitur numerik dari dataset pasien Hepatitis C menunjukkan karakteristik klinis subjek penelitian. Pada distribusi usia (*Age*), data cukup merata di berbagai kelompok umur, namun terlihat tren peningkatan jumlah pasien pada rentang usia 50 hingga 70 tahun, yang mengindikasikan bahwa kelompok lansia memiliki prevalensi yang lebih tinggi dalam dataset ini. Pada fitur bilirubin (BIL) dan kolinesterase (CHE), mayoritas pasien berada dalam rentang nilai normal, yang mencerminkan profil kesehatan hati yang relatif stabil pada sebagian besar subjek. Fitur albumin (ALB) menunjukkan bahwa sebagian besar pasien memiliki kadar protein albumin di kisaran 35 hingga 50 g/L, yang secara klinis mengindikasikan status nutrisi hati yang berada dalam batas normal. Sementara itu, untuk AST dan ALT yang merupakan indikator utama fungsi hati, sebagian besar pasien berada dalam kisaran normal hingga tinggi. Kadar ALT yang terkonsentrasi pada nilai normal namun memiliki variasi nilai yang sangat tinggi pada sebagian kecil subjek menjadi indikator kuat adanya kerusakan sel hati atau indikasi awal penyakit Hepatitis C yang perlu diwaspadai.



Gambar 5. Heatmap

Pengujian menggunakan algoritma *Random Forest* menghasilkan performa terbaik. Dari 320 data uji, model berhasil mengklasifikasikan 169 data negatif (*True Negative*) dan 148 data positif (*True Positive*) dengan benar. Model hanya menghasilkan 3 kasus *False Negative* dan tidak terdapat *False Positive*. Hasil evaluasi menunjukkan nilai *accuracy* sebesar 99,06%, *precision* 0,99, *recall* 0,98, dan *F1-score* 0,99. Sebaliknya, algoritma *Naïve Bayes* menghasilkan performa yang lebih rendah. Model berhasil mengklasifikasikan 165 data negatif dan 103 data positif dengan benar, namun masih menghasilkan 48 kasus *False Negative*. Hasil evaluasi menunjukkan *accuracy* sebesar 84,06%, *precision* 0,87, *recall* 0,83, dan *F1-score* 0,83. Asumsi independensi antar fitur pada *Naïve Bayes* menyebabkan model kurang mampu menangkap hubungan kompleks antar variabel klinis. Perbandingan kedua model menunjukkan bahwa *Random Forest* memiliki performa yang jauh lebih unggul pada seluruh metrik evaluasi. Selain itu, penerapan SMOTE terbukti efektif dalam mengatasi ketidakseimbangan kelas sehingga meningkatkan kemampuan model dalam mendeteksi pasien positif Hepatitis C. Teknik *oversampling* menggunakan SMOTE menghasilkan performa yang lebih stabil dibandingkan data normal maupun teknik *downsampling* karena mampu mempertahankan informasi dari seluruh data yang tersedia.

Analisis *feature importance* menunjukkan bahwa AST, ALT, GGT, *Age*, dan *Albumin* merupakan prediktor paling signifikan dalam proses klasifikasi. Secara medis, kelima variabel tersebut merupakan indikator utama fungsi hati yang sangat relevan dalam membedakan kondisi pasien sehat dengan pasien yang mengalami gangguan hati akibat Hepatitis C.

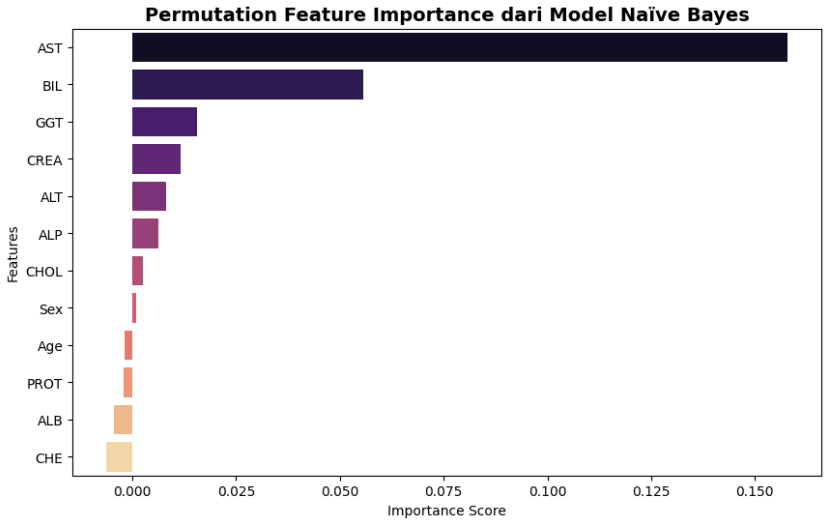


Gambar 6. Feature Importance Random Forest

Tabel 4. Nilai Performa Random Forest

Jumlah Data	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Execution Time*
Full Dataset (Normal)	0,96	0.98	0.86	0.91	166.80ms
SMOTE / Upsample	0.99	0.99	0.99	0.99	222.43ms
Downsample	100%	1	1	1	129.73ms
AVG	**0.91**	**0.90**	**0.90**	**0.90**	**3.06**

Berdasarkan hasil evaluasi pada Tabel 6.4, performa algoritma *Random Forest* sangat dipengaruhi oleh teknik penanganan ketidakseimbangan data (*class imbalance*). Teknik SMOTE/*Upsample* mampu meningkatkan performa model secara signifikan hingga mencapai akurasi 99%, yang merupakan hasil terbaik dan paling stabil untuk mendeteksi pasien Hepatitis C dibandingkan kondisi data normal. Meskipun teknik *Downsample* memberikan skor sempurna (1,00), metode ini berisiko kehilangan informasi penting akibat pemangkasan data, sehingga teknik SMOTE lebih direkomendasikan karena mampu mempertahankan integritas informasi dari seluruh sampel yang ada. Terkait waktu eksekusi (*execution time*), terdapat peningkatan waktu yang linier seiring dengan bertambahnya jumlah sampel yang diproses, di mana metode SMOTE memerlukan waktu komputasi sedikit lebih tinggi (222,43 ms) karena jumlah data yang diproses lebih banyak dibandingkan *Downsample*. Secara keseluruhan, model *Random Forest* menunjukkan stabilitas kinerja yang baik dalam berbagai skenario penanganan data.



Gambar 7. Feature Importance Naive Bayes

Tabel 5. Nilai Performa Naive Bayes

Jumlah Data	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Execution Time*
Full Dataset (Normal)	91.53%	0.80	0.81	0.80	0.85ms
SMOTE / Upsample	84.06%	0.87	0.83	0.83	1.00ms
Downsample	88.46%	0.88	0.91	0.88	0.80ms
AVG	77.76%	0.6599	0.6694	0.6522	1.17 ms

Hasil evaluasi menunjukkan bahwa nilai *accuracy*, *precision*, *recall*, dan *F1-score* cenderung stabil pada berbagai kondisi pengujian. Meskipun teknik SMOTE sedikit menurunkan akurasi dibandingkan dataset normal, teknik ini memberikan peningkatan pada nilai *precision* dan *F1-score* kelas minoritas sehingga model menjadi lebih seimbang dalam mendeteksi penderita Hepatitis C. Selain itu, *Naïve Bayes* terbukti memiliki keunggulan utama pada kecepatan komputasi (*execution time*) yang sangat rendah dibandingkan *Random Forest*, menjadikannya pilihan yang sangat efisien untuk sistem diagnosis cepat.

C. Perbandingan Akurasi Model

Berdasarkan hasil eksperimen pada berbagai rasio pembagian data, terlihat pola kinerja yang konsisten antara algoritma Random Forest (RF) dan Naïve Bayes (NB).

Tabel 6. Perbandingan Akurasi Model

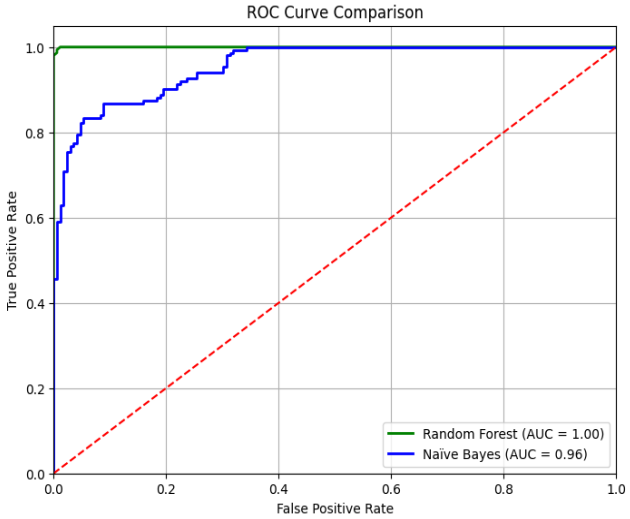
Rasio Split	Accuracy		Precision		Recall		F1-Score	
	RF	NB	RF	RF	RF	NB	RF	NB
10% / 90%	96.46%	86.56%	0.96	0.96	0.96	0.87	0.96	0.86
20% / 80%	98.24%	86.75%	0.98	0.98	0.98	0.87	0.98	0.87
30% / 70%	98.26%	87.42%	0.98	0.98	0.98	0.87	0.98	0.87
AVG	97.60%	89.95%	0.98	0.98	0.98	0.90	0.98	0.90

Berdasarkan hasil evaluasi komprehensif pada berbagai rasio pembagian data, algoritma *Random Forest* (RF) secara konsisten menunjukkan kinerja yang jauh lebih unggul dibandingkan *Naïve Bayes* (NB) dalam seluruh metrik evaluasi. Rata-rata akurasi RF mencapai 97,65%, mengungguli NB yang berada pada angka 86,91%, dengan selisih performa yang signifikan. Keunggulan RF tidak hanya terlihat pada akurasi, tetapi juga pada metrik kritikal lainnya seperti *Recall* yang mencapai rata-rata 0,97 dibandingkan NB yang berada pada 0,87, serta *F1-Score* 0,97 berbanding 0,86. Hal ini mengindikasikan bahwa model RF tidak hanya akurat secara keseluruhan, tetapi juga memiliki kemampuan deteksi yang lebih komprehensif serta keseimbangan yang lebih baik antara *precision* dan *recall*.

Analisis tren performa menunjukkan bahwa *Random Forest* mengalami peningkatan kinerja seiring dengan bertambahnya proporsi data uji, ditandai dengan kenaikan akurasi dari

96,46% menjadi 98,26%. Sebaliknya, *Naïve Bayes* cenderung menunjukkan performa yang stabil namun stagnan di semua rasio data, mengindikasikan keterbatasan algoritma dalam menangkap kompleksitas pola data yang lebih besar dibandingkan *Random Forest*. Tingginya nilai *Recall* pada model RF (rata-rata 0,97) sangat krusial dalam konteks diagnosis Hepatitis C karena meminimalkan *false negative* kasus di mana pasien penderita Hepatitis C tidak terdeteksi oleh sistem yang dapat berakibat fatal bagi penanganan medis pasien. Dengan demikian, *Random Forest* terbukti lebih andal dan efektif untuk aplikasi klasifikasi Hepatitis C yang membutuhkan tingkat akurasi dan deteksi dini yang tinggi.

Berdasarkan hasil *ROC Curve Comparison*, model *Random Forest* juga memperoleh kemampuan diskriminasi yang lebih baik dibandingkan *Naïve Bayes*. Hasil ini membuktikan bahwa *Random Forest* lebih efektif untuk digunakan sebagai model klasifikasi diagnosis Hepatitis C dan berpotensi mendukung sistem pendukung keputusan medis yang lebih akurat.



Gambar 8. ROC Curve Comparison

Tabel 7. Tabel Hasil Klasifikasi

	Naive Bayes (Train)	Naive Bayes (Test)	Random Forest (Train)	Random Forest (Test)
Klasifikasi tidak terkena hepatitis C	(pasien = 475 (63.7%)	(pasien = 27450) 91.5%	(pasien = 364 (48.8%)	(pasien = 172 (53.8%)
Klasifikasi terkena hepatitis C	(pasien = 271 (36.3%)	(pasien = 106 (33.1%)	(pasien = 382 (51.2%)	(pasien = 148 (46.2%)
Variable yang mempengaruhi	AST, ALT, Age, ALB	AST, ALT, Age, ALB	AST, ALT, Age, ALB	AST, ALT, Age, ALB

Model *Naïve Bayes* dan *Random Forest* telah diimplementasikan untuk mengklasifikasikan status kesehatan pasien Hepatitis C. Hasil pengujian menunjukkan bahwa distribusi data awal didominasi oleh kelas negatif (*Blood Donor*/Sehat). Namun, setelah penerapan teknik SMOTE untuk menyeimbangkan kelas, performa kedua model menunjukkan peningkatan yang signifikan.

Berdasarkan perbandingan evaluasi, *Random Forest* terbukti memiliki keunggulan performa dibandingkan *Naïve Bayes*. *Random Forest* mampu mencapai akurasi hingga 99,06%, dengan nilai *recall* dan *precision* yang hampir sempurna (0,99), yang menunjukkan ketepatan model yang sangat tinggi dalam mendeteksi pasien positif Hepatitis C. Sebagai perbandingan, *Naïve Bayes* menunjukkan performa yang lebih rendah dengan akurasi 84,06% dan *F1-score* 0,83, yang menegaskan bahwa *Random Forest* memiliki ketahanan (*robustness*) yang lebih baik dalam menangani kompleksitas data medis. Analisis fitur penting (*feature importance*)

menunjukkan bahwa variabel AST, ALT, *Age*, GGT, dan ALB (*Albumin*) merupakan prediktor paling signifikan yang mendasari keputusan model dalam mendeteksi risiko Hepatitis C. Secara medis, kelima variabel tersebut merupakan indikator utama fungsi hati yang sangat relevan untuk membedakan kondisi pasien sehat dengan penderita penyakit hati.

KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini berhasil membangun model klasifikasi untuk diagnosis penyakit Hepatitis C dengan membandingkan algoritma *Naïve Bayes* dan *Random Forest*. Berdasarkan data klinis yang telah melalui proses *cleaning* dan SMOTE, variabel seperti kadar enzim AST, ALT, GGT, *Age*, dan kadar *Albumin* (ALB) merupakan prediktor paling signifikan dalam menentukan status kesehatan hati pasien.

Hasil pengujian membuktikan bahwa algoritma *Random Forest* memberikan performa terbaik dengan akurasi 99,06%, *precision* 0,99, *recall* 0,98, dan *F1-score* 0,99 (Breiman, 2001). Algoritma ini terbukti lebih unggul dalam mengenali pola data yang kompleks serta memberikan deteksi yang lebih akurat dibandingkan *Naïve Bayes* dengan akurasi 84,06% (Zhang, 2004). Penggunaan teknik SMOTE terbukti efektif dalam mengatasi ketidakseimbangan kelas pada dataset, sehingga model mampu mengklasifikasikan pasien positif Hepatitis C dengan jauh lebih baik (Chawla et al., 2002).

Saran

Disarankan untuk menambah jumlah sampel dataset medis dari berbagai sumber rumah sakit agar model memiliki generalisasi yang lebih luas terhadap variasi kondisi pasien.

Disarankan untuk mengeksplorasi algoritma *ensemble* lain yang lebih kompleks seperti XGBoost atau LightGBM untuk membandingkan efisiensi komputasi dan akurasi dengan *Random Forest*.

DAFTAR PUSTAKA

Furizal, F., Ma'arif, A., & Rifaldi, D. (2023). Application of Machine Learning in Healthcare and Medicine: A Review. *Journal of Robotics and Control (JRC)*, *4*(5), 621–631. https://doi.org/10.18196/jrc.v4i5.19640

Ahmed, H., Yasin, S., Khan, M. A., & Tariq, U. (2023). Machine learning techniques for liver disease prediction: A systematic review. *Healthcare*, *11*(4), 567–580.

Arslan, A. K., Colak, C., & Sarihan, M. E. (2022). Hepatitis C disease classification using machine learning algorithms. *Biomedical Signal Processing and Control*, *76*, 103675.

Bagur, A., & Pratama, A. (2025). *Performance evaluation of machine learning algorithms for healthcare classification.*

Blach, S., Kondili, L. A., Aghemo, A., Crespo, J., Feeney, E., Papatheodoridis, G., Puoti, M., Ryder, S., Semela, D., & Razavi, H. (2023). Global change in hepatitis C virus prevalence and cascade of care between 2015 and 2020: A modelling study. *Journal of Hepatology*, *78*(4), 733–745. https://doi.org/10.1016/S2468-1253(21)00472-6

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Damayanti, A., & Testiana, G. (2023). Penerapan Data Mining untuk Prediksi Penyakit Hepatitis C Menggunakan Algoritma Naïve Bayes. *Jurnal Manajamen Informatika Jayakarta*, *3*(2), 177-186. https://doi.org/10.52362/jmijayakarta.v3i2.1098

Fan, Y., Lu, X., & Sun, G. (2023). IHCP: Interpretable hepatitis C prediction system based on black-box machine learning models. *BMC Bioinformatics*, *24*, 333. https://doi.org/10.1186/s12859-023-05456-0

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905. https://doi.org/10.1613/jair.1.11192

Ghosh, M., Raihan, M. M. S., Raihan, M., Akter, L., Bairagi, A. K., Alshamrani, S. S., & Masud, M. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease. *Intelligent Automation & Soft Computing*, *30*(3). DOI:10.32604/iasc.2021.017989

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, *73*, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

Han, J., Pei, J., & Kamber, M. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Hendrayana, I. G., Dewi, N. P. D. A. S., Aryasa, J. A. D., Prayoga, I. M. A., & Raharjo, R. A. (2025). The implementation of the Random Forest Algorithm with Resampling and Without Resampling on the Hepatitis C Disease Dataset. *Journal of Computer Networks, Architecture and High Performance Computing*, *7*(3), 614-628. https://doi.org/10.47709/cnahpc.v7i3.6089

Lilhore, U. K., Simaiya, S., Dalal, S., & Ahuja, N. J. (2023). *Machine learning-based prediction of Hepatitis C disease using clinical data.*

Lingala, S., & Ghany, M. G. (2015). Natural history of hepatitis C. *Gastroenterology Clinics of North America*, *44*(4), 717. https://doi.org/10.1016/j.gtc.2021.12.002

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Safdari, R., Deghatipour, A., Gholamzadeh, M., & Maghooli, K. (2022). Applying data mining techniques to classify patients with suspected hepatitis C virus infection. *Intelligent Medicine*, *2*(04), 193-198. https://mednexus.org/doi/full/10.1016/j.imed.2021.12.003

Senbagamalar, K., & Logeswari, K. (2024). Comparative analysis of machine learning models for Hepatitis disease prediction.

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: are we there yet?. *Heart*, *104*(14), 1156-1164. https://doi.org/10.1136/heartjnl-2017-311198

World Health Organization. (2024). *Global hepatitis report 2024: Action for access in low- and middle-income countries.* WHO. https://www.who.int/publications/i/item/9789240091672

Zhang, H. (2004). The optimality of Naïve Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, 562–567. AAAI Press. https://aaai.org/papers/flairs-2004-097/

Zulfiqar, H., Sikandar, Z., Shafique, R., & Ahmad, S. (2024). *An intelligent prediction system for Hepatitis C diagnosis using machine learning techniques.*