Vol. 4, No. 3, Desember 2023, hal. 772-777

CoSciTech Computer Science and Information Technology

p-ISSN: 2723-567X
e-ISSN: 2723-5661

Jurnal Computer Science and Information Technology (CoSciTech)

http://ejurnal.umri.ac.id/index.php/coscitech/index

Analisa kinerja algoritma machine learning untuk prediksi virus hepatitis C

Rahmad Gunawan[1], Muhammad Ilham Pratama[2]
Email: [1]goengoen78@umri.ac.id, [2]pratamailham049@gmail.com

[1,2]Teknik Informatika, Ilmu Komputer, Universitas Muhammadiyah Riau

Diterima: 28 Desember 2023 | Direvisi: - | Disetujui: 1 Januari 2024

Abstrak

Virus hepatitis C (HCV) adalah virus RNA dan salah satu patogen manusia yang dilahirkan melalui darah yang disebut sebagai Hepatitis C. Menurut World Health Organization (WHO), diperkirakan hampir 3% atau 120-130 juta penduduk dunia terinfeksi HCV dan 3-4 juta kasus infeksi baru. Diagnosa dini HCV belum efektif sehingga sebagian besar faktor yang berkontribusi terhadap penyakit masih belum jelas. Tujuan penelitian ini adalah mengimplementasikan algoritma machine learning untuk mengidentifikasi faktor yang berkontribusi terhadap virus hepatitis C dan permasalahan prediksi virus hepatitis C dengan melakukan perbandingan pada tiap algoritma untuk menentukan algoritma terbaik dalam memprediksi virus hepatitis C pada dataset *HCV UCI Machine Learning Repository*. Enam algoritma pengklasifikasi di usulkan yaitu *Naive Bayes, Decision Tree, Logictic Regression, K-Nearest Neighbor, Support Vector Machine,* dan *Random Forest.* Hasil menunjukkan dari nilai akurasi tiap algoritma didapati algoritma terbaik untuk memprediksi virus hepatitis C yaitu *random forest* dengan tingkat akurasi 98.37% dan didapati fitur yang paling berkontribusi terhapat model prediksi pasien infected HCV dan not infected HCV yaitu AST(*Aspartat aminotransferase*) dan ALP(*alkaline phosphatase*).

Kata kunci: *Virus Hepatitis C, Prediksi, Machine Learning, Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Random Forest*

Machine Learning Algorithm Performance Analysis For Hepatitis C Virus Prediction

Abstract

Hepatitis C (HCV) is an RNA virus and one of the blood-borne human pathogens known as Hepatitis C. According to the World Health Organization (WHO), it is estimated that nearly 3% or 120-130 million of the world's population are infected with HCV and 3-4 million new infection cases. Early diagnosis of HCV has not been effective so most of the factors that contribute to the disease are still unclear. This study aims to implement a machine learning algorithm to identify factors that contribute to hepatitis C virus and hepatitis C virus prediction problems by comparing each algorithm to determine the best algorithm for predicting hepatitis C virus in the HCV UCI Machine Learning Repository dataset. Six classification algorithms are proposed: Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Random Forest. The results show that from the accuracy value of each algorithm, the best algorithm for predicting hepatitis C virus is random forest with an accuracy rate of 98.37% and it was found that the features that contributed the most to the prediction model for HCV-infected and non-HCV patients were AST (Aspartate aminotransferase) and ALP (alkaline phosphatase).

Keywords: *Hepatitis C Virus, Prediction, Machine Learning, Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Random Forest*

1. PENDAHULUAN

Virus hepatitis C (HCV) adalah virus RNA dan salah satu patogen manusia yang dilahirkan melalui darah yang disebut sebagai Hepatitis C [1]. Menurut *World Health Organization* (WHO), diperkirakan hampir 3% atau 120-130 juta penduduk dunia

terinfeksi HCV dan 3-4 juta kasus infeksi baru[2]. HCV sejenis virus yang ditularkan melalui darah. Bentuk infeksi yang umum terjadi pada sejumlah kecil darah. Penularan dapat terjadi karena penggunaan obat suntik, perawatan kesehatan yang tidak aman, transfusi darah atau plasma darah, praktik injeksi yang tidak aman, dan praktik seksual[3]. HCV kronis dapat merusak hati secara perlahan. Sayangnya pasien mungkin tidak menunjukkan gejala sampai penyakit tersebut berkembang menjadi siriosis hati [4]

Diagnosa dini HCV belum efektif sehingga sebagian besar faktor yang berkontribusi terhadap penyakit masih belum jelas[4]. Oleh karena itu mengembangkan teknik machine learning untuk dapat mengidentifikasi dan memprediksi apakah pasien tersebut terinfeksi HCV ataupun tidak terinfeksi mungkin dapat bermanfaat lebih untuk menangani penyakit tersebut, dengan mempertimbangkan beberapa faktor yang mengidentifikasi penyakit tersebut seperti *Albumin, Bilirubin, Kolin Esterase, Glutamyl Transferase, Aspartat Aminotransferase, Alanine Amino Transferase, Kolesterol, Kreatinin , Protein , dan Alkaline Phosphatase* [5].

Machine learning merupakan salah satu cara efektif yang dapat digunakan dalam dunia biomedis[6]. Hal ini terlihat dari bagaimana *machine learning* mendekati pendekatan pembuatan algoritma yang baik dan otomatis yang dapat digunakan dalam proses diagnosa atau prediksi penyakit untuk proses pengambilan keputusan melihat jumlah data yang dihasilkan di bidang kesehatan, dan juga sulitnya proses pengelolaan data, berbagai pendekatan ditawarkan dengan menggunakan metode *machine learning* yang ada [7].

Penelitian-penelitian sebelumnya menunjukkan perkembangan prediksi algoritma *machine learning* pada penyakit HCV seperti penelitian yang dilakukan oleh Ashfaq Ali Kashif dkk (2021) hasil penelitian tersebut menunjukan kinerja terbaik didapat oleh *K-Nearest Neighbor* dan *random forest* dengan tingkat akurasi 89.52% [8], penelitian oleh Hiroaki Haga dkk (2020) peneltian tersebut menggunakan beberapa algortima *machine* mendapatkan tingkat akurasi cukup baik diantaranya *Support Vector Machine* 93.7%, *Logictic Regression* 84.8% dan *Decision Tree* 74.0%[9] dan penelitian oleh Sneha Grampurohit dkk (2020) hasil penelitian tersebut menunjukan kinerja terbaik didapat oleh *naïve bayes* dengan tingkat akurasi 93.61%[10].

Penelitian ini bertujuan menganalisa kinerja algoritma untuk memprediksi keputusan diagnosa virus hepatitis C dengan menerapkan *machine learning*, seperti pada penelitian-penelitian sebelumnya yang menggunakan beberapa *algoritma machine learning* medapatkan hasil yang baik. Penelitian ini melalukan klasifikasi pasien terinfeksi virus hepatitis C dan pasien yang tidak terinfeksi virus hepatitis C dengan menggunakan beberapa algoritma *machine learning* yaitu *Naive Bayes, Decision Tree, Logictic Regression, K-Nearest Neighbor, Support Vector Machine,* dan *Random Forest.* Data yang digunakan dalam penelitian ini adalah dataset virus hepatitis C yang bersumber dari *UCI Machine learning Respository*, dataset tersebut mengandung 14 fitur atau fitur termasuk fitur class [11].

2. METODE PENELITIAN

2.1 Indentifikasi Dataset

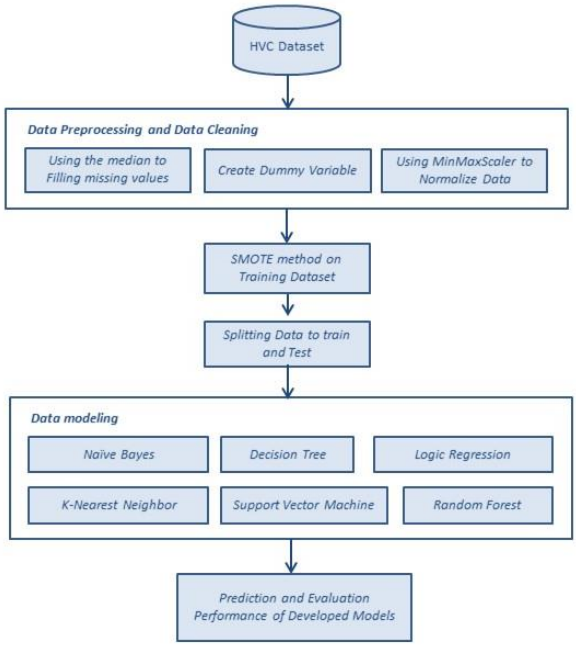
Dataset yang digunakan adalah *public* dataset yang diambil dari https://archive.ics.uci.edu/ml/datasets/HCV+data. Pada dataset (Tabel 1) yang akan di gunakan yaitu dataset Hepatitis C Virus *UCI Machine Learning Respository* nantinya akan dilakukan penambahan fitur berupa kategori pasien.

Tabel. 1 Penjelasan Dataset HCV

Variable	***Description***	***Type***	***Range***
Age	Usia pasien	*Integer*	19–77
Sex	Laki-laki dan Perempuan	*Categorical*	F dan M
ALB	Mengukur jumlah albumin dalam darah	*Real*	14.9–82.2
ALP	*alkaline phosphatase* enzim dalam aliran darah yang bertugas membantu memecah protein dalam tubuh. Organ hati menjadi salah satu sumber utama ALP	*Real*	11.3–416.6
ALT	*Aanine aminotransferase*, menunjukkan kerusakan hati akibat hepatitis, infeksi, sirosis, kanker hati, atau penyakit hati lainnya	*Real*	0.9–325.3
AST	*Aspartat aminotransferase* adalah enzim yang terkait dengan kinerja organ hati	*Real*	10.6–324
BIL	Tes bilirubin mengukur jumlah bilirubin dalam darah	*Real*	0.8–254
CHE	*Kolinesterase serum* adalah enzim yang disintesis oleh hepatosit dan kadarnya mencerminkan fungsi sintetik hati	*Real*	1.42–16.41
CHOL	Mengukur jumlah kolesterol dan *trigliserida* dalam darah	*Real*	1.43–9.67
CREA	Tes kreatinin adalah ukuran seberapa baik ginjal melakukan tugasnya menyaring limbah dari darah	*Real*	8–107.9

CGT	*Gamma glutamyl transferase* adalah tes fungsi hati yang bertujuan untuk menilai kondisi kesehatan organ hati	*Real*	4.5–650
PROT	Tes protein total mengukur jumlah protein dalam darah	*Real*	44.8–90

2.2 Mengusulkan Arsitektur Prediksi Virus Hepatitis C Dengan Algoritma *Machine Learning*



Gambar 1. Arsitektur dari Metode yang Diusulkan

Adapun tahapan preprosessing sebagai berikut:
Data Preprocessing dan Data Cleaning

1. *Filling Missing Value*
   Dalam proses ini melakukan pengecekan missing value pada dataset kemudian menggunakan median untuk mengganti value yang hilang, value yang hilang tersebut diganti dengan nilai median dari seluruh kolom fitur.
2. *Dummy Variable*
   Dalam proses ini melakukan pengecekan missing value pada dataset kemudian menggunakan median untuk mengganti value yang hilang, value yang hilang tersebut diganti dengan nilai median dari seluruh kolom fitur.Dalam proses membuat dummy variable melakukan konversi data kategori ke bentuk numerik yang terdiri dari dua nilai yaitu 0 dan 1.
3. *MinMaxScaler to Normalize Data*
   Dalam proses ini melakukan normalisasi data menggunakan fitur class dari sklearn yaitu MinMaxScaler agar fitur memiliki rentang nilai yang sama, tidak ada yang terlalu besar maupun terlalu kecil sehingga dapat membuat analisis data menjadi lebih mudah.
4. Smote, dataset terdiri dari kelas data yang tidak seimbang, pemisahan pelatihan secara acak, dan satu set tes menghasilkan hasil kelas yang tidak merata. Proses ini menerapkan teknik smote untuk menyeimbangkan data menggunakan imblearn library.
5. Splitting data, setelah di lakukan proses penyeimbangan data maka di peroleh dataset yang baru, yang mana dataset yang baru ini nanti nya akan displit atau dibagi menjadi dua yaitu: data training (80%) dan data testing (20%) yang nantinya di gunakan dalam pengklasifikasian.
6. Data modeling, data yang dibagi menjadi dua, pada bagian *data training* (80%) di terapkan ke *algoritma naïve bayes*, *decision tree*, *logic regression*, *K-Nearest neighbor*, *Support vector machine*, dan *random forest* untuk mengetahui tingkat akurasi pada data yang digunakan. Dengan menggunakan parameter yang berkaitan dengan model.
7. Evaluasi kinerja, mengevaluasi hasil kinerja pada tiap algoritma, dalam penelitian ini teknik evaluasi model algoritma yang digunakan adalah *confution matrix*. *Confution matrix* adalah cara termudah untuk mengevaluasi hasil kinerja algoritma dengan membandingkan beberapa contoh positif yang diklasifikasikan benar atau salah, dan negatif yang diklasifikasikan benar atau salah. *Confusion matrix* mempunyai pandangan berangam, dimana *confusion matrix* mempunyai peran mendasar dalam mengevaluasi kinerja algoritma klasifikas, Dalam *confusion matrix*, baris merepresentasikan label yang benar, dan kolom merepresenatsikan hasil prediksi algoritma pengklasifikasi, Tabel 2 menunjukkan *confution matrix performance*.

Tabel 2. *Convution Matrix Evaluasi Performance*

	Prediksi Invected HCV	**Prediksi Not Infected HCV**
Invected HCV	TP	TN
Not Infected HCV	FP	FN

1. TP (*True Positif*) merupakan data positif infected HCV yang di prediksi benar oleh algoritma pengklasifikasi.
2. TN (*True Negatif*) merupakan data positif infected HCV yang di prediksi salah oleh algoritma pengklasifikasi.
3. FP (*False Positif*) merupakan data not infected HCV dan algortima pengklasifikasi memprediksi dengan benar.
4. FN (*False Negatif*) merupakan data not infected HCV dan algortima pengklasifikasi memprediksi dengan salah.

Kinerja model yang dikembangkan dibandingkan dengan cara menghitung *Accuracy*, *recall (sensitivity)*, *specificity*, *precision*, dan *F-measure* (*F1 score*) pada tiap algoritma. Berikut adalah persamaan untuk menghitung *Accuracy*, *recall (sensitivity)*, *specificity*, *precision*, dan *F-measure* (*F1 score*) :

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{..... (1)}$$

$$Specitify = \frac{TN}{TN+FP} \quad \text{..... (2)}$$

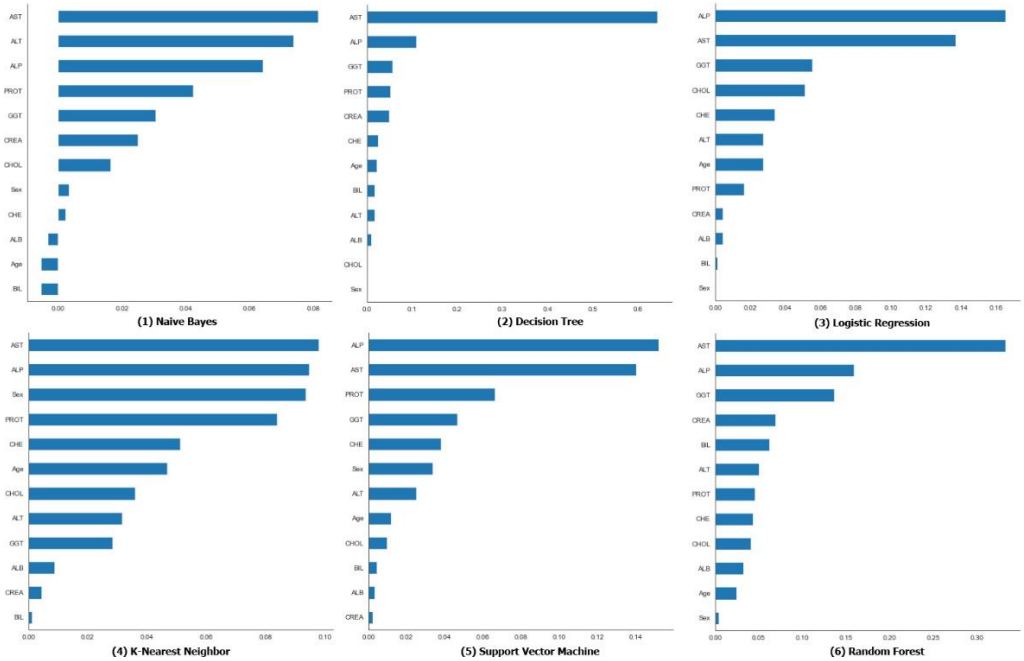
$$Recall = \frac{TP}{TP+FP} \quad \text{..... (3)}$$

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad \text{..... (4)}$$

Setelah mendapatkan hasil *Accuracy*, *recall (sensitivity)*, *specificity*, *precision*, dan *F-measure* (*F1 score*) dari tiap algoritma, selanjutnya melihat variable yang paling berpengaruh terhadap model prediski dengan feature important sehingga bisa mendapatkan tolak ukur besaran kontribusi berbagai variable data yang dilatih kepada kinerja model prediks. Kemudian hasil dari tiap algoritma akan dibandingkan menentukan kinerja algoritma terbaik untuk prediksi virus hepatitis C.

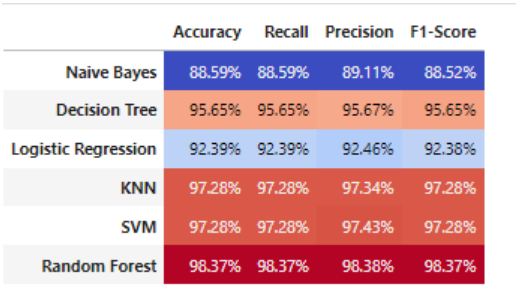
3. HASIL DAN PEMBAHASAN

Tahap modeling pada dataset HCV akan melakukan pemodelan data dengan beberapa algoritma *machine learning* yang digunakan pada penelitian ini (Gambar 2), adapun algoritma yang akan digunakan untuk pemodelan dataset HCV adalah *naïve bayes*, *decision tree*, *logic regression*, *K-Nearest neighbor*, *Support vector machine*, dan *random forest*. Tahapan ini mengklasifikasikan dataset HCV menjadi 2 kategori yaitu infected HCV dan not infected HCV kemudian data dilatih dengan rasio data *train* 80% dan data *test* 20% pada tiap algoritma yang digunakan pada penelitian dan pada tahapan ini juga melihat fitur pada dataset yang berkontribusi terhadap model dengan *feature important* (Gambar 2).



Gambar 2. Feature Important Tiap Model Prediksi

Tahapan ini mengevaluasi hasil kinerja pada tiap algortima yang digunakan pada penelitian ini dengan melihat nilai prediksi pada tiap algortima menggunakan confusion matrix, melihat hasil *score accuracy*, *precision*, *recall*, dan *f-score* dari hasil *confusion matrix* kemudian mengevaluasi fitur yang berkontribusi pada tiap model prediksi dengan *feature important*.



	Accuracy	Recall	Precision	F1-Score
Naive Bayes	88.59%	88.59%	89.11%	88.52%
Decision Tree	95.65%	95.65%	95.67%	95.65%
Logistic Regression	92.39%	92.39%	92.46%	92.38%
KNN	97.28%	97.28%	97.34%	97.28%
SVM	97.28%	97.28%	97.43%	97.28%
Random Forest	98.37%	98.37%	98.38%	98.37%

Gambar 3 Perbandingan *Accuracy Model* Prediksi

Dapat dilihat pada gambar 3 hasil accuracy tiap algoritma *machine learning* yang digunakan pada penelitian ini, urutan pertama didapat oleh *random forest* dengan *score* 98.37%, kedua didapat oleh *support vector machine* dengan *score* 97.28%, ketiga didapat oleh k-*nearest neighbor* dengan *score* 97.28%, keempat didapat oleh decision tree dengan *score* 95.56%, kelima didapat oleh logistic regression dengan score 92.39%, urutan terakhir didapat oleh *naïve bayes* dengan *score* 88.59%. Selanjutnya perbandingan hasil *confusion matrix* pada tiap algoritma, berikut tabel 3 perbandingan hasil *confusion matrix* pada tiap algoritma :

Tabel 3. Perbandingan Hasil *Confusion Matrix*

Algoritma		**Label**	**Prediksi**	
			Infected HCV	Not Infected HCV
Naïve Bayes	**Aktual**	Infected HCV	90	5
		Not Infected HCV	16	73
Decision Tree		Infected HCV	92	3
		Not Infected HCV	5	84
Logistic Regression		Infected HCV	90	5
		Not Infected HCV	9	80
K-Nearest Neighbor		Infected HCV	91	4
		Not Infected HCV	1	88
Support Vector Machine		Infected HCV	90	5
		Not Infected HCV	0	89
Random Forest		Infected HCV	93	2
		Not Infected HCV	1	88

Dari table 3 diatas berisikan hasil *confusion matrix* tiap algoritma yang digunakan pada penelitian ini, dapat dilihat algoritma *random forest* mendapat nilai *false positif* (FP) dan *false negative* (FN) paling rendah dari tiap algoritma. Nilai *false positif* (FP) dan *false negative* (FN) merupakan kesalahan prediksi pada model yang mana jika pasien tidak terinfeksi virus hepatitis C tetapi diprediksi terinfeksi virus hepatitis C (FP), maka pada diagnosa selanjutnya pasien tersebut dapat mengetahui keadaan sebenarnya bahwa pasien tersebut benar tidak terinfeksi virus hepatitis C. Tetapi jika ada pasien yang sebenarnya terinfeksi virus hepatitis C tetapi diprediksi tidak terinfeksi virus hepatitis C (FN), maka pasien tersebut akan mengetahui keadaan sebenarnya dengan sangat terlambat dan pasien tersebut tidak segera mengambil tindakan pencegahan medis untuk hepatitis C itu. Sehingga dapat menyebabkan kondisi pasien yang semakin memburuk hingga menyebabkan seriosis atau fibrosis hati bahkan kematian.

8. KESIMPULAN

Berdasarkan hasil evaluasi kinerja tiap *algortima machine* learning pada dataset HCV dari *UCI repository* yang telah dibahas pada pembahasan sebelumnya, penelitian menghasilkan output berupa *score accuracy* dan hasil prediksi *confusion matrix* dari tiap algortima kemudian fitur yang berkontribusi pada tiap model prediksi yaitu *feature important*. Penelitian ini telah mencapai hasil akhir dan didapatkan kesimpulan sebagai berikut :

1. Kinerja dari tiap algoritma *machine learning* yang diterapkan pada dataset HCV untuk pemodelan prediksi virus hepatitis C, telah mendapatkan hasil *accuracy* dari tiap algrotima, *naïve bayes* mendapatkan akurasi 88.59%, *decision tree* mendapatkan akurasi 95.56%, *logistic regression* mendapatkan akurasi 92.39%, *k-nearest neighbor* mendapatkan akurasi 97.27%, *support vector machine* mendapat akurasi 97.27%, dan *random forest* mendapat akurasi 98.37%. dari nilai akurasi tiap algoritma didapati algoritma terbaik untuk memprediksi virus hepatitis C yaitu *random forest* dengan tingkat akurasi 98.37%.
2. Dari tiap algoritma yang dilatih untuk pemodelan prediksi virus hepatitis C didapati fitur yang paling berkontribusi terhapat model prediksi pasien infected HCV dan not infected HCV yaitu AST (*Aspartat aminotransferase*) merupakan enzim yang terkait dengan kinerja organ hati dan ALP (*alkaline phosphatase*) enzim dalam aliran darah yang bertugas membantu

memecah protein dalam tubuh, organ hati menjadi salah satu sumber utama dari ALP yang mana penyakit hepatitis C itu sendiri adalah penyakit peradangan pada hati.

DAFTAR PUSTAKA

[1] A. Petruzziello, S. Marigliano, G. Loquercio, A. Cozzolino, and C. Cacciapuoti, "Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes," *World J. Gastroenterol.*, vol. 22, no. 34, pp. 7824–7840, 2016, doi: 10.3748/wjg.v22.i34.7824.

[2] L. Syafa'ah, Z. Zulfatman, I. Pakaya, and M. Lestandy, "Comparison of Machine Learning Classification Methods in Hepatitis C Virus," *J. Online Inform.*, vol. 6, no. 1, p. 73, 2021, doi: 10.15575/join.v6i1.719.

[3] M. B. Butt *et al.*, "Diagnosing the Stage of Hepatitis C Using Machine Learning," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/8062410.

[4] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intell. Med.*, no. September 2021, 2022, doi: 10.1016/j.imed.2021.12.003.

[5] F. Mostafa, E. Hasan, M. Williamson, and H. Khan, "Statistical Machine Learning Approaches to Liver Disease Prediction," *Livers*, vol. 1, no. 4, pp. 294–312, 2021, doi: 10.3390/livers1040023.

[6] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Comput. Sci.*, vol. 162, no. Itqm 2019, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.

[7] G. V. Nivaan and A. W. R. Emanuel, "Analytic Predictive of Hepatitis using the Regression Logic Algorithm," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020*, pp. 106–110, 2020, doi: 10.1109/ISRITI51436.2020.9315365.

[8] A. A. Kashif, B. Bakhtawar, A. Akhtar, S. Akhtar, N. Aziz, and M. S. Javeid, "Treatment Response Prediction in Hepatitis C Patients using Machine Learning Techniques," *Int. J. Technol. Innov. Manag.*, vol. 1, no. 2, pp. 79–89, 2021, doi: 10.54489/ijtim.v1i2.24.

[9] H. Haga *et al.*, "A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus," *PLoS One*, vol. 15, no. 11 November, pp. 1–12, 2020, doi: 10.1371/journal.pone.0242028.

[10] S. Grampurohit and C. Sagarnal, "Disease prediction using machine learning algorithms," *2020 Int. Conf. Emerg. Technol. INCET 2020*, pp. 1–7, 2020, doi: 10.1109/INCET49848.2020.9154130.

[11] Lichtinghagen, R., Klawonn, F., Hoffmann, G. 2020. HCV data Data Set [online]. Tersedia : https://archive.ics.uci.edu/ml/datasets/HCV+data. Diakses pada 26 mai 2022.