Klasifikasi Penyakit Hepatitis C dengan Menggunakan *K-Nearest Neighbor*

Fathul Qorib Yusfila[1*], Bain Khusnul Khotimah[2]
Devie Rosa Anamisa[3], Ana Tsalitsatun Ni'mah[4]
[1,2,3] Teknik Informatika, Universitas Trunojoyo Madura, Indonesia
[4] Pendidikan Informatika, Universitas Trunojoyo Madura, Indonesia
Jl. Raya Telang, PO BOX 2, Kamal, Bangkalan - 69162
E-mail: 190411100041@student.trunojoyo.ac.id , bain@trunojoyo.ac.id ,
devros_gress@trunojoyo.ac.id, ana.tsalits@trunojoyo.ac.id

DOI : https://doi.org/10.52620/sainsdata.v3i1.205

ABSTRAK

Hepatitis merupakan kondisi peradangan pada hati yang disebabkan oleh berbagai jenis virus, baik yang menular maupun tidak menular, dan dapat menimbulkan komplikasi serius hingga kematian. Terdapat lima tipe utama virus hepatitis, yaitu A, B, C, D, dan E. Penelitian ini bertujuan untuk mengklasifikasikan penyakit Hepatitis C menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan pendekatan penanganan data tidak seimbang melalui *teknik Random Oversampling*. Dataset yang digunakan adalah HCV dari UCI *Machine Learning Repository*, yang terdiri dari 615 data dengan 14 fitur dan 5 kategori kelas. Karena data bersifat tidak seimbang, dilakukan peningkatan jumlah data pada kelas minoritas menggunakan *Random Oversampling*. Proses evaluasi dilakukan dengan membandingkan performa KNN tanpa dan dengan *oversampling*, serta menentukan nilai K terbaik melalui skenario pengujian menggunakan 5-*fold Cross Validation*. Hasil menunjukkan bahwa KNN tanpa *oversampling* menghasilkan akurasi tertinggi sebesar 94% pada nilai K=3, sementara dengan *oversampling* akurasi meningkat menjadi 96,70% pada nilai K yang sama. Dengan demikian, dapat disimpulkan bahwa penerapan *Random Oversampling* mampu meningkatkan performa klasifikasi algoritma KNN pada data Hepatitis C yang tidak seimbang.

Kata Kunci: Hepatitis C, *K-Nearest Neighbor,* klasifikasi*, Random Oversampling*, *Cross Validation.*

ABSTRACT

Hepatitis is a liver inflammation condition caused by various types of viruses, both infectious and non-infectious, which can lead to serious complications and even death. There are five main types of hepatitis viruses: A, B, C, D, and E. This study aims to classify Hepatitis C using the K-Nearest Neighbor (KNN) algorithm with a handling approach for imbalanced data through the Random Oversampling technique. The dataset used is the HCV dataset from the UCI Machine Learning Repository, consisting of 615 records with 14 features and 5 class categories. Due to the imbalance in the data, the minority classes were increased using Random Oversampling. The evaluation process was carried out by comparing the performance of KNN with and without oversampling, as well as determining the optimal value of K through test scenarios using 5-fold Cross Validation. The results show that KNN without oversampling achieved the highest accuracy of 94% at K=3, while with oversampling, the accuracy increased to 96.70% at the same K value. Thus, it can be concluded that the application of Random Oversampling enhances the classification performance of the KNN algorithm on imbalanced Hepatitis C data.

Keywords: Hepatitis C, K-Nearest Neighbor, Classification, Random Oversampling, Cross Validation.

PENDAHULUAN

Hepatitis merupakan penyakit peradangan pada hati yang disebabkan oleh berbagai jenis virus, baik yang bersifat menular maupun tidak menular, yang dapat menimbulkan masalah kesehatan serius hingga berakibat fatal. Terdapat lima tipe utama virus hepatitis, yaitu A, B, C, D, dan E. Di antara kelima tipe tersebut, virus hepatitis B dan C berpotensi menyebabkan sirosis dan kanker hati, yang menjadi penyakit kronis pada ratusan juta orang di dunia dan merupakan salah satu penyebab utama kematian. Saat ini, tercatat sekitar 354 juta orang di dunia hidup dengan infeksi hepatitis B atau C [1].

Hepatitis C terutama ditularkan melalui jalur parenteral, seperti penggunaan jarum suntik secara bergantian atau melalui transfusi darah. Sebaliknya, penularan melalui hubungan seksual tergolong jarang terjadi [2]. Salah satu upaya penanganan terhadap infeksi virus hepatitis C adalah dengan terapi antivirus langsung (*Direct-Acting Antiviral* / DAA). Terapi ini menunjukkan tingkat kesembuhan yang tinggi dan memiliki tingkat toleransi yang baik. DAA terdiri dari kombinasi dua atau lebih obat yang dikonsumsi secara oral, sehingga lebih praktis dibandingkan terapi injeksi. Penggunaan DAA mempertimbangkan genotipe dan kondisi sirosis pasien, serta memerlukan perhatian khusus untuk penderita hepatitis C dengan komorbid seperti HIV atau hepatitis B, pasien yang menjalani transplantasi organ dan jaringan, anak-anak, serta wanita hamil [3].

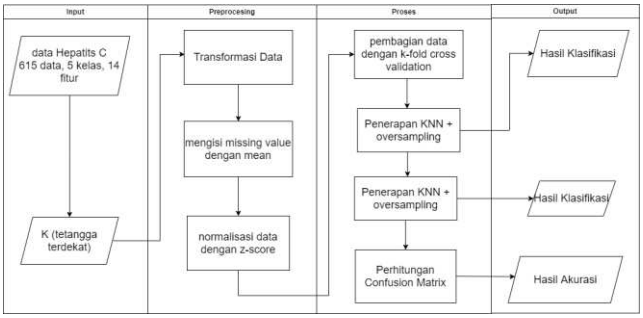
Seiring berkembangnya teknologi, pemanfaatan *data mining* menjadi penting dalam mendukung proses diagnosis dan klasifikasi penyakit. *Data mining* merupakan proses ekstraksi informasi yang tersembunyi dan potensial dari sejumlah besar data, serupa dengan proses analisis yang dilakukan oleh analis data [4][5][6]. Salah satu metode dalam *data mining* adalah klasifikasi, yaitu proses membangun model untuk memetakan data ke dalam kelas-kelas yang telah ditentukan sebelumnya[7]. Beberapa algoritma yang umum digunakan untuk klasifikasi antara lain *Naive Bayes*, *Support Vector Machine*, dan *K-Nearest Neighbor* (KNN). Algoritma KNN bekerja dengan cara mencari sejumlah K data terdekat dari data baru yang akan diklasifikasikan, lalu menentukan kelas data tersebut berdasarkan mayoritas tetangga terdekatnya [8][9][10].

Beberapa penelitian sebelumnya telah membandingkan performa algoritma KNN dengan algoritma lainnya. Penelitian oleh [11] menunjukkan bahwa algoritma *Gaussian Naive Bayes* menghasilkan akurasi sebesar 90,98%, presisi 69,91%, dan *recall* 61,57%, sementara KNN menunjukkan akurasi yang lebih tinggi yaitu 91,80%, dengan presisi 68,96% dan *recall* 51,85%. Penelitian lain oleh [12] menyimpulkan bahwa baik *Naive Bayes* maupun KNN mampu mengklasifikasikan penyakit hati dengan baik. Dalam penelitian tersebut, *Naive Bayes* mencatat akurasi, presisi, dan *recall* sebesar 85,5%, sedangkan KNN menunjukkan hasil sempurna dengan akurasi, presisi, dan *recall* masing-masing sebesar 100%.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan performa algoritma KNN dalam mengklasifikasikan penyakit Hepatitis C sebelum dan sesudah dilakukan proses *resampling* menggunakan teknik *Random Oversampling*. Dataset yang digunakan merupakan data multivariat dari UCI Machine Learning Repository, yang terdiri atas 14 fitur dan 5 kategori kelas. Penelitian ini dilakukan dengan menggunakan bahasa pemrograman Python, pengukuran performa model menggunakan *Confusion Matrix*, serta validasi model melalui metode *K-Fold Cross Validation*.

METODE

Rancangan sistem klasifikasi Hepatitis C yang digunakan pada penelitian ini yaitu digambarkan dalam bentuk IPO (*Input,Process,Output*) yang ditunjukan pada gambar 1.



Gambar 1. Rancangan Sistem Klasifikasi Hepatitis C

Preprocessing

Preprocessing pada penelitian ini terdapat 3 tahap yaitu:

1. *Data Transformation*
   Data Transformation atau transformasi data adalah metode yang digunakan untuk merubah data dengan tipe kategorikal menjadi tipe data numerik[13].
2. *Mean Imputation*
   Mean Imputation atau imputasi rata rata adalah Teknik pengisian data yang hilang dengan menghitung rata rata nilai dari variabel yang sama. rumus *Mean Imputation* tesebut yaitu.

$$\bar{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} \quad (1)$$

Keterangan:

$\bar{x}$	= mean
$x_1 + x_2 + x_3 + \cdots + x_n$	= jumlah seluruh data nilai
n	= jumlah seluruh frekuensi

3. *Z-Score Normalization*
   Data kemudian di normalisasi menggunakan *Z-Score Normalization* di mana konsep dari normalisasi ini yang berdasarkan nilai *mean* atau nilai rata-rata dan *Standart Deviation* (deviasi standart) dari data. Metode ini sangat berguna jika tidak diketahui nilai aktualisasi maksimum dan minimum pada data[14]. Rumus *Z-Score Normalization* sebagai berikut:

$$X_{new} = \frac{X-\mu}{\sigma} \text{ atau } = \frac{X-Mean(X)}{StDev(x)} \quad (2)$$

Keterangan :

X_{new}	= nilai baru dari normalisai data
X	= nilai lama
μ	= nilai populasi *mean*
σ	= nilai standart deviasi

Random Oversampling

Random Oversampling (ROS) adalah metode *resampling* kelas minoritas dalam data secara acak, kemudian data yang dipilih diduplikasi dan ditambahkan ke dataset *training* baru sehingga jumlah kelas minoritas sama dengan kelas mayoritas[15].

K-Nearest Neighbor

K-Nearest Neighbor (KNN) melakukan klasifikasi data objek berdasarkan jumlah K dari data pelatihan terdekat. Tujuan dari klasifikasi ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut data sampel yang ada dalam data *Training*. Dalam metode KNN, langkah-langkah yang dilakukan adalah menentukan parameter K atau jumlah tetangga terdekat, menghitung jarak antara pasangan data menggunakan metrik jarak *Euclidean Distance.* Setelah jarak dihitung, hasil pengukuran jarak diurutkan secara menaik dari nilai jarak terkecil hingga terbesar. Langkah selanjutnya adalah mengelompokkan berdasarkan target klasifikasi dan menentukan kelas dengan menggunakan nilai K yang telah ditentukan sebelumnya. Kelas akan ditentukan berdasarkan mayoritas nilai K yang akan menentukan kelas tersebut [16].
Adapun Langkah Langkah pada algoritma KNN sebagai berikut:

1. Tentukan K untuk jumlah tetangga.
2. Hitung jarak menggunakan *Euclidean Distance* antara data *Training* dan data *Testing*.
   Jarak *Euclidean* dapat dilihat dari persamaan berikut

$$d = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (3)$$

3. Urutkan jarak dan indeks yang terurut dari terkecil ke terbesar berdasarkan perhitungan jarak.
4. Tentukan K dari data terdekat.
   Tentukan kelompok *Testing* berdasarkan label mayoritas pada K.

K-Fold Cross Validation

K-fold Cross Validation adalah salah satu bentuk *Cross Validation* yang sering digunakan. Metode ini membagi data menjadi k subset, kemudian dilakukan proses pelatihan sebanyak k kali. Pada setiap iterasi, k-1 subset digunakan sebagai data pelatihan, sementara subset yang tersisa digunakan sebagai data validasi. [17].

Confusion Matrix

Confusion Matrix adalah sebuah tabel klasifikasi yang berisi informasi hasil perhitungan secara keseluruhan. Tabel ini digunakan untuk mengevaluasi pengukuran melalui akurasi, presisi, dan recall. Untuk memudahkan pembacaan, hasil evaluasi tersebut direpresentasikan dalam bentuk tabel klasifikasi.[6].

Tabel 1. Confusion Matrix

		Nilai sebenarnya	
		Benar	Salah
Nilai Prediksi	Positif	TP (benar positif)	FP (salah positif)
	Negatif	TN (benar negatif)	FN (salah negatif)

Pada *Confusion Matrix* terdapat 4 istilah sebagai representasi hasil proses klasifikasi yaitu *True Positive* (TP), *False Positive* (FP), *True Negatif* (TN) dan *False Negatif* (FN)[5]. Akurasi mengukur persentase hasil klasifikasi sistem yang tepat. Presisi adalah ukuran akurasi untuk kelas tertentu, sedangkan recall mengindikasikan persentase data yang diklasifikasikan sebagai positif dari keseluruhan data yang sebenarnya positif [13]. Adapun rumus nya sebagai berikut:

$$\text{Akurasi} = \frac{TP}{TP+TN+FP+FN} x\ 100\% \quad (4)$$

$$\text{Presisi} = \frac{TP}{TP+FP} x100\% \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} x100\% \quad (6)$$

Di mana :
a). TP "*True Positive*", yaitu jumlah data positif yang diklasifikasikan dengan benar oleh sistem.
b). TN "*True Negatif*", yaitu jumlah data negatif yang diklasifikasikan dengan benar oleh sistem.
c). FP "*False Positive*", yaitu jumlah data positif tetapi diklasifikasikan salah oleh sistem.
d). FN "*False Negatif*", yaitu jumlah data negatif tetapi diklasifikasikan salah oleh sistem.

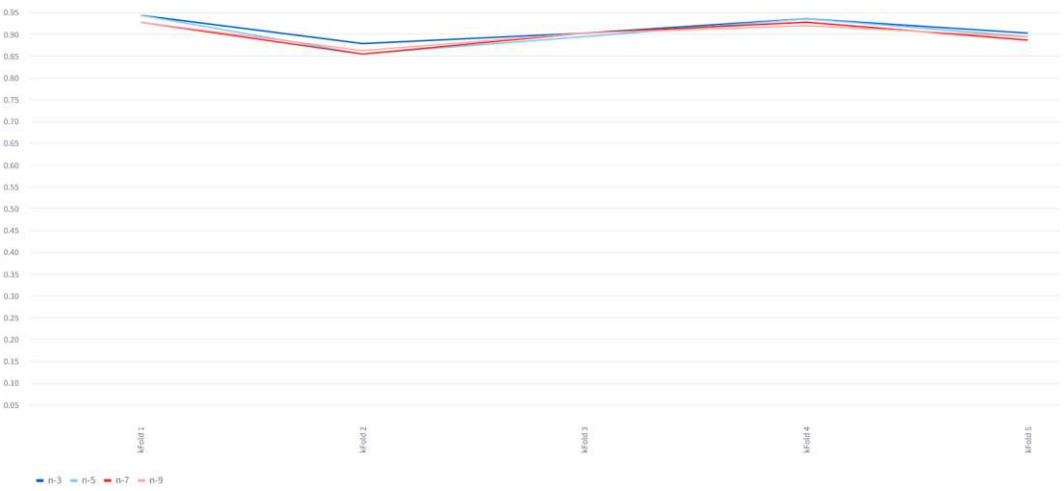
HASIL DAN PEMBAHASAN

Perbandingan hasil akurasi dari percobaan K-nearest = 3,5,7, dan 9 pada KNN tanpa menggunakan oversampling dapat dilihat pada tabel 2 dibawah ini.

Tabel 2. Perbandingan K-nearest = 3,5,7, dan 9 pada KNN

K-nearest	Metode	Model Evaluasi			Prediksi	Kelas prediksi	Waktu (s)
		Akurasi	Presisi	Recall			
3	KNN	94.30%	96.46%	99.09%	109	0=Blood Donor	0.5644
			50%	50%	2	1=Hepatitis	
			100%	25%	1	2=Fibrosis	
			80%	80%	4	3=Cirrhosis	
5	KNN	94.30%	95.65%	100%	110	0=Blood Donor	0.5549
			50%	25%	1	1=Hepatitis	
			66%	50%	2	2=Fibrosis	
			100%	60%	3	3=Cirrhosis	
7	KNN	92.68%	94.02%	100%	107	0=Blood Donor	0.4658
			0%	0%	3	1=Hepatitis	
			50%	25%	3	2=Fibrosis	
			100%	60%	4	3=Cirrhosis	
9	KNN	92.68%	93.22%	100%	107	0=Blood Donor	0.3913
			0%	0%	3	1=Hepatitis	
			50%	25%	3	2=Fibrosis	
			100%	60%	3	3=Cirrhosis	

Hasil akurasi dari KNN dengan K-nearest = 3,5,7, dan 9 dengan pembagian data menggunakan 5-*fold Cross Validation* dapat dilihat pada gambar 2 dibawah ini.



Gambar 2. Grafik rata-rata akurasi pada KNN dengan 5-fold Cross Validation

Grafik diatas jika disimpulkan berdasarkan hasil rata rata akurasi dapat dilihat pada tabel 3 dibawah ini.

Tabel 1. Rata-rata akurasi KNN

K-fold	K-nearest-3	K-nearest-5	K-nearest-7	K-nearest-9	Rata-rata
k-fold 1	94%	94%	93%	93%	93%
k-fold 2	88%	85%	85%	86%	86%
k-fold 3	90%	89%	90%	90%	90%
k-fold 4	94%	94%	93%	92%	93%
k-fold 5	90%	89%	89%	89%	89%

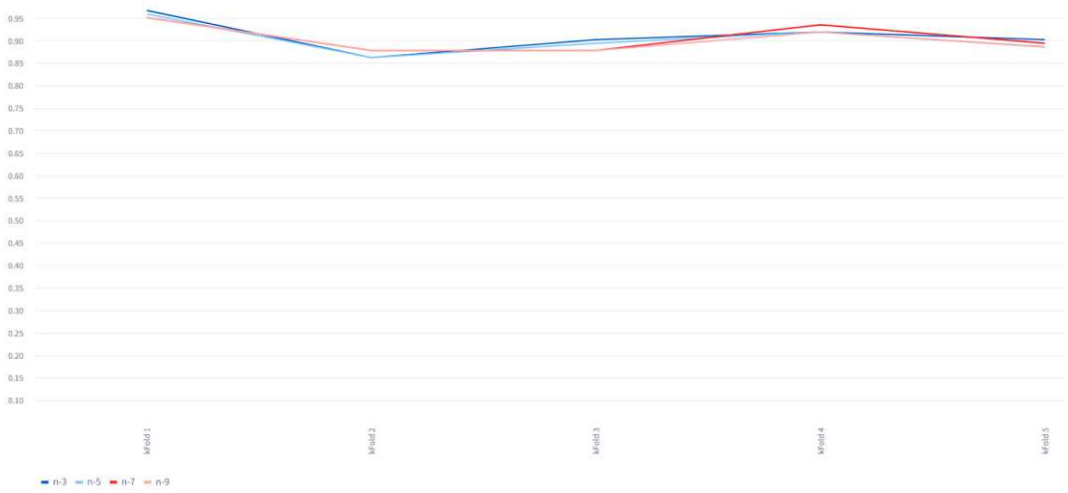
Pada tabel diatas dapat dilihat bahwa pembagian data dengan *5-fold Cross Validation* diperoleh hasil rata-rata akurasi model yang berbeda. Pada k-fold = 1, model memperoleh rata-rata akurasi tertinggi sebesar 93%, menunjukan kinerja yang sangat baik. Namun, akurasi tersebut mengalami penurunan pada k-fold = 2 dengan nilai sebesar 86%. Peningkatan kembali terjadi pada k-fold = 3 dengan nilai rata-rata akurasi sebesar 90%, menunjukan pembagian dataset menjadi tiga subset memberikan dampak kontribusi positif terhadap performa model. Selanjutnya k-fold = 4, k-fold = 5 menunjukan akurasi sebesar 93% dan 89% secara berturut-turut. Meskipun terdapat kenaikan pada k-fold = 4, hasil ini menunjukan bahwa model memiliki kemampuan yang baik untuk menggeneralisasi pada data yang tidak terlihat selama proses *training.*

Perbandingan hasil akurasi dari percobaan K-nearest = 3,5,7, dan 9 pada KNN menggunakan *oversampling* dapat dilihat pada tabel 4.10 dibawah ini.

Tabel 4. Perbandingan K-nearest = 3,5,7, dan 9 pada KNN + *oversampling*

K-nearest	Metode	Model Evaluasi			Prediksi	Kelas prediksi	Waktu (s)
		Akurasi	Presisi	Recall			
3	KNN + *oversampling*	96.70 %	99.09% 60% 100% 80%	99.09% 75% 75% 80%	109 3 3 4	0=Blood Donor 1=Hepatitis 2=Fibrosis 3=Cirrhosis	1.9795
5	KNN + *oversampling*	95.90 %	100% 50% 100% 66.67%	98.19% 75% 75% 80%	108 3 3 4	0=Blood Donor 1=Hepatitis 2=Fibrosis 3=Cirrhosis	2.0422
7	KNN + *oversampling*	95.12 %	100% 42.86% 100% 66.67%	97.27% 75% 75% 80%	107 3 3 4	0=Blood Donor 1=Hepatitis 2=Fibrosis 3=Cirrhosis	1.9868
9	KNN + *oversampling*	95.12 %	100% 42.86% 100% 66.67%	97.27% 75% 75% 80%	107 3 3 4	0=Blood Donor 1=Hepatitis 2=Fibrosis 3=Cirrhosis	2.0008

Hasil akurasi dari KNN + *oversampling* dengan K-nearest = 3,5,7, dan 9 dengan pembagian data menggunakan 5-*fold Cross Validation* dapat dilihat pada gambar 3 dibawah ini.



Gambar 3. Grafik rata-rata akurasi pada KNN + oversampling dengan 5-fold Cross Validation

Grafik diatas jika disimpulkan berdasarkan hasil rata rata akurasi dapat dilihat pada tabel 5 dibawah ini.

Tabel 5. rata-rata KNN+oversampling

k-fold	K-nearest-3	K-nearest-5	K-nearest-7	K-nearest-9	Rata-rata
k-fold 1	97%	96%	95%	95%	96%
k-fold 2	86%	86%	88%	88%	87%
k-fold 3	90%	89%	88%	88%	89%
k-fold 4	92%	92%	94%	92%	92%
k-fold 5	90%	89%	89%	89%	89%

Pada tabel diatas dapat dilihat bahwa pembagian data dengan *5-fold Cross Validation* diperoleh hasil rata-rata akurasi model yang bervariasi. Pada k-fold = 1 model mencapai akurasi tertinggi dengan nilai 96%, menunjukan kemampuan model untuk memberikan prediksi yang tepat jika diuji pada subset tertentu. Namun, pada k-fold = 2 mengalami penurunan akurasi pada model sebesar 87%, yang mungkin disebabkan oleh variasi data yang kurang representatif dalam dua subset. Peningkatan kembali terjadi pada k-fold = 3 dengan rata-rata akurasi sebesar 89%, menunjukan bahwa pembagian dataset menjadi 3 subset dapat meningkatkan generalisasi model. Ketika k-fold = 4 rata-rata akurasi mengalami peningkatan kembali menjadi 92% yang menandakan bahwa lebih banyak subset dalam proses *cross validation* dapat meningkatkan kinerja model. Pada k-fold = 5 rata-rata akurasi mencapai 89% menunjukan hasil yag baik meskipun mengalami penurunan akurasi.

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, diperoleh perbandingan performa algoritma *K-Nearest Neighbor* (KNN) dalam mengklasifikasikan data Hepatitis C sebelum dan sesudah diterapkan metode *Random Oversampling*. KNN tanpa oversampling menghasilkan akurasi tertinggi sebesar 94% pada nilai *K-nearest* = 3. Setelah diterapkan *Random Oversampling*, akurasi meningkat menjadi 97% pada nilai *K-nearest* yang sama. Peningkatan akurasi sebesar 3% ini menunjukkan bahwa metode *oversampling* efektif dalam menangani ketidakseimbangan data (*imbalanced data*) dan meningkatkan kinerja model klasifikasi. Selain itu, variasi akurasi yang diperoleh dari setiap pengujian menunjukkan bahwa pembagian data menggunakan *k-fold cross validation* turut memengaruhi hasil klasifikasi. Oleh karena itu, pemilihan nilai *k-fold* yang tepat sangat penting untuk memperoleh hasil yang optimal dan representatif.

Saran

Meskipun hasil klasifikasi menggunakan algoritma KNN menunjukkan performa yang cukup baik, penelitian ini masih memiliki keterbatasan, salah satunya adalah dominasi satu kelas dalam hasil klasifikasi, baik sebelum maupun sesudah diterapkannya metode *oversampling*. Hal ini kemungkinan besar disebabkan oleh proses pembagian data secara acak melalui *k-fold cross validation* serta pemilihan metode *resampling* yang belum sepenuhnya sesuai dengan karakteristik data. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi metode pembagian data dan teknik *resampling* lain yang lebih sesuai, seperti *SMOTE (Synthetic Minority Over-sampling Technique)* atau *stratified sampling*, guna mengurangi dominasi kelas dan meningkatkan generalisasi model. Selain itu, dapat pula dilakukan perbandingan dengan algoritma klasifikasi lain seperti *Random Forest*, *Support Vector Machine*, atau *XGBoost* untuk memperoleh wawasan yang lebih komprehensif terkait performa klasifikasi data Hepatitis C.

DAFTAR PUSTAKA

[1] Aditya, A., Mustofa, F. L., Hidayat, H., & Firlanda, Z. R. S. (2022). Prevalensi Hepatitis C Pada Donor Darah Sebelum Dan Pada Saat Pandemi Covid 19 Di Unit Transfusi Darah Palang Merah Indonesia Provinsi Lampung Tahun 2019-2020. Malahayati Nursing Journal, 4(6), 1544-1556. https://doi.org/10.33024/mnj.v4i6.6460

[2] Kurniawati, S. A., Karjadi, T. H., & Gani, R. A. (2015). Faktor-faktor yang berhubungan dengan kejadian hepatitis C pada pasangan seksual pasien koinfeksi Human Immunodeficiency Virus dan virus hepatitis C. Jurnal Penyakit Dalam Indonesia, 2(3), 133-139.

[3] Saraswati, A. *et al.*, "Faktor Risiko Terjadinya Penyakit Hepatitis C," 2022, [Online]. Available: http://jurnal.globalhealthsciencegroup.com/index.php/JPPP

[4] Senduk, V. Y., & Parmadi, E. H. (2022). Klasifikasi User yang Berpotensi Melakukan Pembelian Barang Online Menggunakan Algoritme Weighted K-Nearest Neighbor. In Prosiding Seminar Nasional Ilmu Sosial dan Teknologi (SNISTEK). 4 pp. 109-114.

[5] Setiawan, S. B., Adiwijaya, A., & Mubarok, M. S. (2018). Klasifikasi Topik Berita Berbahasa Indonesia menggunakan Weighted K-Nearest Neighbor. eProceedings of Engineering, 5(1).

[6] Ni'mah, A. T., & Syuhada, F. (2022). Term Weighting Based Indexing Class and Indexing Short Document for Indonesian Thesis Title Classification. *Journal of Computer Science and Informatics Engineering (J-Cosine)*, *6*(2), 167-175. https://doi.org/10.29303/jcosine.v6i2.471

[7] Khotimah, B. K., & Syarief, M. (2011). Aplikasi Data Mining Untuk Mengukur Tingkat Kelulusan Mahasiswa Dengan Metode Apriori Dan K-Mean Clustering (Studi Kasus Jurusan Teknik Informatika UTM). Jurnal Simantec, 2(2), 71-80. https://doi.org/10.21107/simantec.v2i2.13394

[8] Toyibah, Z. B., Putri, Y. N., Puandini, P., Widodo, Z. M., & Ni'mah, A. T. (2024). Perbandingan Kinerja Algoritma Multinomial Naïve Bayes dan Logistic Regression pada Analisis Sentimen Movie Ratings IMDB. *Jurnal Ilmiah Edutic: Pendidikan dan Informatika*, *10*(2), 181-189.

[9] Ni'mah, A. T., & Yunitarini, R. (2024). Relevance of the Retrieval of Hadith Information (RoHI) using Bidirectional Encoder Representations from Transformers (BERT) in religious education media. In *BIO Web of Conferences* (Vol. 146, p. 01041). EDP Sciences. https://doi.org/10.1051/bioconf/202414601041

[10] Ni'mah, A. T., & Arifin, A. Z. (2020). Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis. Rekayasa, 13(2), 172–180. https://doi.org/10.21107/rekayasa.v13i2.6412

[11] Raharja, K. Y., Oktavianto, H., & Umilasari, R. (2021). Perbandingan Kinerja Algoritma Gaussian Naive Bayes Dan K-Nearest Neighbor (KNN) Untuk Mengklasifikasi Penyakit Hepatitis C Virus (HCV). 1-12.

[12] Desiani, A. (2022). Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati. Jurnal Sistem Informasi Dan Sistem Komputer, 7(2), 104-110. https://doi.org/10.51717/simkom.v7i2.96

[13] Dafwen, T. (2021). Optimasi Nilai k Pada Algoritma k Nearest Neighbor Untuk Prediksi Akademik Mahasiswa Yang Bekerja. Indonesian Journal of Computer Science, 10(2), 379-388. https://doi.org/10.33022/ijcs.v10i2.3005

[14] Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan normalisasi data untuk klasifikasi wine menggunakan algoritma K-NN. Comput. Eng. Sci. Syst. J, 4(1), 78. https://doi.org/10.24114/cess.v4i1.11458

[15] Batan, G. A., Keytimu, M. J., Katumbo, F. L., Binanto, I., & Sianipar, N. F.(2023) Penerapan Metode Random Forest, Gaussian NB, Dan KNN Terhadap Data Unbalance dan Data Balance Menggunakan Random Over Sampling Untuk Klasifikasi Senyawa. Keladi Tikus. 8th Seminar Nasional Teknik Elektro, Informatika dan Sistem Informasi (SINTaKS). https://doi.org/10.35842/sintaks.v2i1.26

[16] Yunani, R. (2022) *Pengenalan Tulisan Aksara Lampung Dengan Library Opencv Menggunakan Metode Projection Profile Dan Klasifikasi K-NN*.

[17] Lutfi, M., & Hasyim, M. (2019). Penanganan data missing value pada kualitas produksi jagung dengan menggunakan metode K-NN Imputation pada algoritma C4. 5. Jurnal RESISTOR (Rekayasa Sistem Komputer), 2(2), 89-104.. http://jurnal.stiki-indonesia.ac.id/index.php/jurnalresistor