Prosiding Annual Research Seminar 2025
Computer Science and ICT
ISBN :
Vol.5 No.1

ANALISIS KOMPARATIF MODEL KLASIFIKASI MACHINE LEARNING DENGAN IMPUTASI KNN UNTUK PREDIKSI PENYAKIT HATI KRONIS

Aisyah Nur Khoirofiq
Fakultas Ilmu Komputer
Universitas Sriwijaya
Sumatera Selatan, Indonesia

Annisa Darmawahyuni
Fakultas Ilmu Komputer
Universitas Sriwijaya
Sumatera Selatan, Indonesia

Abstrak*—Data medis yang tidak lengkap dapat menghambat analisis dan menurunkan akurasi model prediktif. Penelitian ini menangani missing value pada dataset sirosis hati menggunakan metode K-Nearest Neighbor (KNN) Imputation. Tantangan utamanya adalah menentukan parameter k yang optimal, yang dipengaruhi oleh bias-variance trade-off. Untuk mengatasinya, penelitian menerapkan pendekatan sistematis dengan GridSearchCV untuk mengoptimalkan parameter k pada KNN Imputer secara simultan dengan hiperparameter tiga model klasifikasi: KNN, SVM, dan Random Forest. Hasil evaluasi menunjukkan bahwa kombinasi KNN Imputation dengan model SVM memberikan kinerja terbaik, mencapai akurasi 73.81% pada data uji. Temuan kunci menyatakan bahwa nilai k optimal untuk imputasi bervariasi tergantung model klasifikasi yang digunakan, sehingga optimasi terintegrasi sangat penting. Namun, semua model masih mengalami keterbatasan dalam memprediksi kelas minoritas akibat ketidakseimbangan data yang ekstrem.

Keyword*—Imputasi Data, KNN Imputation, Penyakit Hati Kronis, Machine Learning, Support Vector Machine (SVM), Ketidakseimbangan Kelas.

I. PENDAHULUAN

Penyakit hati kronis merupakan tantangan kesehatan global yang kian meningkat, termasuk di Indonesia. Berdasarkan data dari Kementerian Kesehatan Republik Indonesia, antara 3,5 hingga 5,2 juta penderita hepatitis B di Indonesia berisiko berkembang menjadi sirosis, yaitu fase akhir dari berbagai kondisi penyakit hati kronis [1]. Sirosis hati ditandai oleh kerusakan pada struktur dan fungsi hati, yang dapat memicu komplikasi serius seperti kanker hati [2]. Oleh karena itu, pemahaman yang mendalam mengenai penyakit ini serta pengelolaan data medis yang tepat sangat penting untuk meningkatkan mutu layanan kesehatan bagi pasien.

Namun, sebuah tantangan umum dalam analisis data medis adalah adanya nilai yang hilang (missing value), yang dapat disebabkan oleh berbagai faktor seperti kesalahan pencatatan atau kegagalan pengukuran. Kehadiran missing value dapat menyebabkan bias dalam analisis dan menurunkan kekuatan statistik dari model prediktif [3, 13]. Untuk mengatasi masalah ini, teknik imputasi data digunakan untuk mengisi nilai yang hilang. Salah satu metode non-parametrik yang populer adalah K-Nearest Neighbor (KNN) Imputation [18, 19]. Metode ini mengestimasi nilai yang hilang berdasarkan nilai dari k tetangga terdekat yang memiliki karakteristik serupa. Keunggulan utama KNN Imputation adalah kemampuannya beradaptasi dengan struktur data lokal yang kompleks dan meminimalkan distorsi pada distribusi variabel yang diimputasi [4].

Pemilihan nilai k merupakan tantangan fundamental yang secara langsung dipengaruhi oleh prinsip bias-variance trade-off [14, 15]. Menurut James et al. (2013), nilai k yang terlalu kecil akan membuat model sangat sensitif terhadap noise (varians tinggi), sementara nilai k yang terlalu besar dapat mengabaikan pola lokal dan menyebabkan model terlalu umum (bias tinggi) [15]. Meskipun terdapat heuristik umum seperti "aturan praktis" k= $\sqrt{N}$ (dimana N adalah jumlah sampel) [14], pendekatan ini tidak selalu optimal. Meskipun beberapa penelitian telah berhasil menerapkan KNN Imputation untuk data medis, banyak di antaranya cenderung menetapkan nilai k untuk imputasi secara terpisah dari proses pemodelan. Masih terdapat celah penelitian (research gap) dalam studi yang secara sistematis mengevaluasi dampak dari optimasi nilai k pada imputer yang dilakukan secara simultan dengan optimasi hiperparameter pada model klasifikasi yang mengikutinya. Interaksi antara kualitas data hasil imputasi dan sensitivitas model klasifikasi—khususnya pada data sirosis hati yang kompleks—belum banyak dieksplorasi.

Oleh karena itu, penelitian ini dirancang untuk secara komprehensif mengisi celah penelitian tersebut. Fokus utamanya adalah menerapkan metode K-Nearest Neighbor (KNN) Imputation pada dataset medis hati kronis, dan kemudian melakukan evaluasi yang sistematis dan kuantitatif terhadap dampaknya pada kinerja tiga model klasifikasi machine learning yang fundamental seperti K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Random Forest. Penelitian ini bergerak melampaui pendekatan konvensional yang sering kali menangani imputasi dan klasifikasi sebagai dua tahap yang terpisah atau modular. Pendekatan terpisah semacam itu berisiko menghasilkan optimasi yang sub-optimal, karena nilai k terbaik untuk imputasi mungkin tidak selaras dengan kebutuhan spesifik dari algoritma klasifikasi yang akan menggunakan data tersebut.

Sebagai gantinya, peneliti ingin mengadopsi sebuah kerangka kerja yang lebih canggih dan terintegrasi dengan memanfaatkan Pipeline dari pustaka Scikit-learn yang dioptimalkan melalui metode GridSearchCV. Pendekatan holistik ini memungkinkan proses pencarian

hiperparameter yang dilakukan secara simultan, di mana nilai k yang paling efektif untuk tahap imputasi dicari secara bersamaan dengan penyesuaian hiperparameter dari setiap model klasifikasi.

Dengan demikian, peneliti tidak hanya bertujuan untuk menemukan model klasifikasi dengan akurasi tertinggi, tetapi juga untuk memberikan bukti empiris yang kuat mengenai sinergi antara tahap pra-pemrosesan data dan pemodelan. Tujuannya adalah untuk mendapatkan pemahaman yang lebih mendalam tentang bagaimana kualitas data yang diimputasi secara langsung memengaruhi konstruksi batas keputusan oleh algoritma yang berbeda, yang pada akhirnya akan menghasilkan model prediktif yang lebih robust dan andal.

II. METODOLOGI PENELITIAN

A. Dataset dan Sumber Data

Dataset yang digunakan dalam penelitian ini adalah Cirrhosis Patient Survival Prediction Dataset dari UCI Machine Learning Repository. Dataset ini berisi data rekam medis dari 418 pasien penderita sirosis hati yang dikumpulkan dari Mayo Clinic antara tahun 1974 hingga 1984. Dataset ini memiliki 20 fitur klinis yang relevan untuk diagnosis dan prognosis penyakit.

Variabel independen (fitur) yang digunakan mencakup berbagai aspek, di antaranya:

- Data Demografis: Usia dalam hari (Age) dan jenis kelamin (Sex).
- Parameter Medis Klinis: Kadar Bilirubin, Kolesterol, Albumin, Tembaga, Alk_Phos, SGOT, Trigliserida, Platelets, dan Prothrombin.
- Data Kondisi Pasien: Adanya Ascites, Hepatomegaly, Spiders, dan Edema.

Variabel dependen (target) yang digunakan dalam analisis ini adalah status kelangsungan hidup pasien (Status), yang dikategorikan ke dalam tiga kelas:

- Kategori C: Censored (pasien masih hidup atau statusnya tidak diketahui pada akhir masa tindak lanjut).
- Kategori D: Death (pasien meninggal dunia).
- Kategori CL: Censored due to Liver Transplant (pasien menjalani transplantasi hati).

B. Pra-pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk menangani nilai yang hilang (missing value) yang terdapat dalam dataset. Pendekatan yang berbeda digunakan untuk tipe data yang berbeda.

- Imputasi Data Kategorikal: Untuk fitur dengan tipe data kategorikal, nilai yang hilang diisi menggunakan modus (mode), yaitu nilai yang paling sering muncul pada fitur tersebut.
- Imputasi Data Numerik: Untuk fitur dengan tipe data numerik, nilai yang hilang diisi menggunakan metode K-Nearest Neighbor (KNN) Imputation. Metode ini mengestimasi nilai yang hilang dengan menghitung rata-rata dari k tetangga terdekatnya. Nilai k yang optimal tidak ditentukan secara manual, melainkan dicari secara sistematis melalui proses optimasi.

C. Model Klasifikasi dan Optimasi Hiperparameter

i. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) adalah metode klasifikasi non-parametrik yang menentukan kelas sebuah data poin baru berdasarkan mayoritas kelas dari k tetangga terdekatnya. Kedekatan antar data diukur menggunakan metrik jarak, salah satu yang paling umum adalah Jarak Euclidean [5], seperti yang ditunjukkan pada Persamaan. Kelemahan utama dari KNN adalah sensitivitasnya terhadap pemilihan nilai k [14]. Nilai k yang terlalu kecil dapat menyebabkan overfitting dan rentan terhadap noise, sementara nilai k yang terlalu besar dapat menyebabkan underfitting dengan mengabaikan pola lokal data

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

dengan:
d = Jarak antara data x dan y
x_i= Fitur ke-i dari data x
y_i= Fitur ke-i dari data y
n = Jumlah fitur

ii. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah model klasifikasi yang bekerja dengan mencari hyperplane atau batas keputusan optimal yang secara maksimal memisahkan data poin dari kelas yang berbeda dalam ruang fitur [15]. SVM efektif di ruang berdimensi tinggi dan fleksibel karena dapat menggunakan berbagai fungsi kernel untuk menangani data yang tidak dapat dipisahkan secara linear.

iii. Random Forest

Random Forest adalah sebuah metode ensemble learning yang membangun banyak pohon keputusan (decision trees) selama proses pelatihan. Untuk klasifikasi, hasil akhir ditentukan oleh suara mayoritas (modus) dari semua pohon individu. Metode ini sangat efektif untuk mengurangi overfitting dan dapat menangani interaksi kompleks antar fitur [15].

iv. Optimasi Hiperparameter dengan GridSearchCV

Untuk menemukan kombinasi hiperparameter terbaik bagi imputer dan classifier secara bersamaan, penelitian ini mengadopsi GridSearchCV (Grid Search with Cross-Validation). Teknik ini secara sistematis menguji semua kombinasi hiperparameter yang telah ditentukan dalam sebuah "grid" dan menggunakan validasi silang 5-lipatan (5-fold cross-validation) untuk mengevaluasi setiap kombinasi secara robust [15]. Pendekatan ini diterapkan

pada sebuah Pipeline yang mengintegrasikan tahap imputasi dan klasifikasi, sehingga memungkinkan penemuan sinergi terbaik antara kedua proses tersebut. Pendekatan validasi silang ini merupakan metode standar untuk mengevaluasi kinerja sebuah classifier [21].

D. Metrik Evaluasi Kinerja

Kinerja model klasifikasi dievaluasi menggunakan Confusion Matrix, yaitu sebuah tabel yang merangkum hasil prediksi dengan membandingkannya terhadap kelas aktual. Matriks ini berisi empat nilai utama: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).

- True Positive (TP): Jumlah data positif yang diprediksi dengan benar.
- True Negative (TN): Jumlah data negatif yang diprediksi dengan benar.
- False Positive (FP): Jumlah data negatif yang salah diprediksi sebagai positif.
- False Negative (FN): Jumlah data positif yang salah diprediksi sebagai negatif.

Berdasarkan nilai-nilai tersebut, dihitung beberapa metrik evaluasi untuk menilai kinerja model secara komprehensif.

E. Confusion Matrix

Sebuah tabel yang merangkum hasil prediksi dengan membandingkannya dengan kelas aktual. Tabel ini berisi empat nilai utama: True Positives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN).

- **Akurasi**—Proporsi total prediksi yang benar ((TP + TN) / Total). Metrik ini memberikan gambaran umum kinerja tetapi bisa menyesatkan pada dataset yang tidak seimbang.
- **Precision**—Kemampuan model untuk tidak melabeli sampel negatif sebagai positif (TP / (TP + FP)).
- **Recall**—Kemampuan model untuk menemukan semua sampel positif (TP / (TP + FN)).
- **F1-Score**—Rata-rata harmonik dari presisi dan recall (2 * (Precision * Recall) / (Precision + Recall)).

Metrik ini sangat berguna ketika terdapat ketidakseimbangan kelas. Metrik-metrik ini adalah dari *Confusion Matrix* seperti Akurasi, Presisi, Recall, dan F1-Score untuk menilai kinerja model secara komprehensif [6].

III. HASIL ANALISIS

A. Skenario Eksperimen

Seluruh eksperimen diimplementasikan menggunakan bahasa pemrograman Python dengan pustaka utama Scikit-learn untuk pemodelan dan Pandas untuk manipulasi data. Dataset dibagi menjadi 80% data latih dan 20% data uji. Proses optimasi hiperparameter untuk KNNImputer dan ketiga model klasifikasi (KNN, SVM, Random Forest) dilakukan secara simultan menggunakan GridSearchCV dengan validasi silang 5-lipatan pada data latih. Model dengan performa terbaik pada validasi silang kemudian dievaluasi secara final pada data uji.

B. Hasil Kinerja Model Klasifikasi

Berikut adalah rangkuman kinerja dari ketiga model pada data uji.

Tabel 1: Perbandingan Kinerja Model pada Data Uji

Model	Akurasi (Data Uji)	Weighted F1-Score	Hiperparameter Optimal Ditemukan
KNN	67.86%	0.65	imputer_k=11, classifier_k=9
SVM	73.81%	0.71	imputer_k=11, C=1, gamma=0.1
Random Forest	72.62%	0.70	imputer_k=19, n_estimators=50, max_depth=5

Berdasarkan Tabel 3.1, model Support Vector Machine (SVM) menunjukkan kinerja terbaik pada data uji dengan akurasi 73.81% dan Weighted F1-Score sebesar 0.71. Kemudian disusul oleh model Random Forest dengan akurasi sebesar 72.62% dan di posisi terkahir dengan model KNN sebesar 67.86%.

Hasil penelitian menunjukkan bahwa model Support Vector Machine (SVM) memberikan kinerja terbaik dengan akurasi 73.81%. Keunggulan ini menegaskan bahwa setelah data dilengkapi dengan KNN Imputation, SVM mampu menemukan batas keputusan yang lebih efektif dalam ruang fitur data medis yang kompleks.

Temuan bahwa KNN Imputation menjadi dasar yang kuat untuk model klasifikasi didukung oleh studi lain. Sebagai contoh, penelitian oleh Pamuji et al. (2024) juga menemukan bahwa KNN Imputation lebih unggul dibandingkan metode sederhana seperti mean imputation pada dataset medis berskala kecil [7]. Li et al. (2024) juga menemukan bahwa KNN Imputation mencapai kinerja terbaik (MAE dan RMSE terendah) dibandingkan tujuh metode lainnya, termasuk mean imputation, regresi, dan MICE [18]. Demikian pula, efektivitas KNN dalam konteks penyakit kronis, meskipun sebagai metode klasifikasi, juga ditunjukkan oleh Tampubolon & Amin (2024) pada kasus diabetes [8]. Peningkatan performa klasifikasi setelah penerapan imputasi juga dilaporkan oleh Arifianto et al. (2022) dan Widyananda et al. (2022) [17, 20].

Salah satu temuan kunci dari penelitian ini adalah nilai k optimal untuk KNNImputer yang bervariasi tergantung pada model klasifikasi yang digunakan k=11 untuk SVM, k=19 untuk Random Forest. Ini menyoroti pentingnya pendekatan optimasi terintegrasi, sejalan dengan penelitian oleh Zhang et al. (2020) yang menekankan bahwa penentuan nilai k yang tepat sangat berpengaruh dalam meningkatkan akurasi [9]. Temuan ini

juga sejalan dengan penelitian oleh Akbar & Kusumodestoni (2020) yang menunjukkan bahwa nilai k terbaik tidak selalu mengikuti kaidah umum, melainkan perlu ditemukan melalui uji coba [16].

C. Analisis Metrik Evaluasi dan Confusion Matrix

i. K-Nearest Neighbors (KNN)

Confusion matrix untuk model KNN disajikan pada Gambar 2. Model ini menunjukkan kesulitan dalam membedakan kelas 'D' (Death), dengan jumlah False Positive yang signifikan (16 sampel).

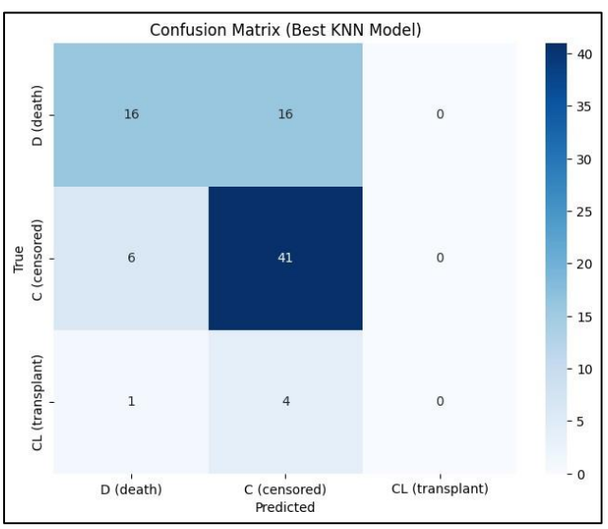

Table 1. Confusion Matrix Best Model KNN

ii. Support Vector Machine (SVM)

Model SVM menunjukkan performa yang lebih seimbang, terutama dalam memprediksi kelas 'D', seperti yang terlihat pada confusion matrix di Gambar 3.

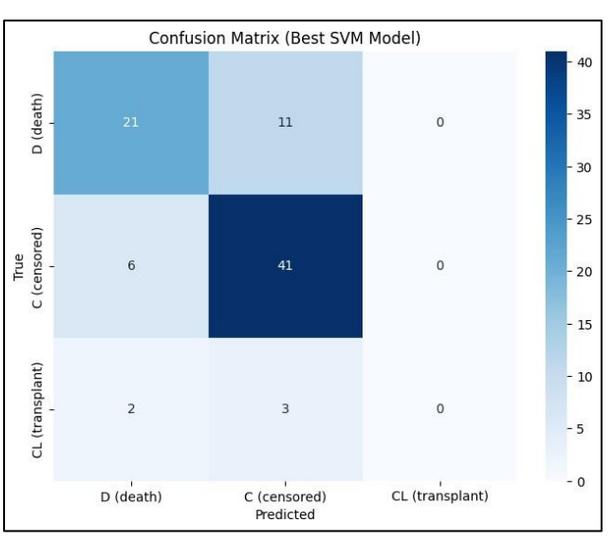

Table 2. Confusion Matrix Best Model SVM

iii. Random Forest

Kinerja Random Forest (Gambar 4) mirip dengan SVM, namun dengan recall yang sedikit lebih rendah untuk kelas 'D'.

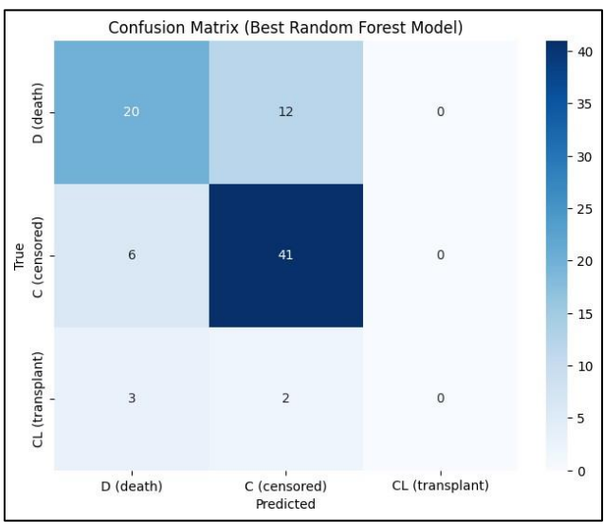

Table 3. Confusion Matrix Best Model Random Forest

Matriks konfusi dari semua model menunjukkan bahwa tidak ada satu pun sampel dari kelas ini yang diprediksi dengan benar. Hal ini disebabkan oleh ketidakseimbangan kelas yang parah dalam dataset, dimana jumlah sampel kelas 'CL' sangat sedikit dibandingkan dengan kelas lainnya. Akibatnya, model belajar bahwa mengabaikan kelas minoritas adalah strategi yang efektif untuk memaksimalkan akurasi secara keseluruhan, yang menyoroti keterbatasan utama dari eksperimen ini.

D. Analisis Pengaruh Parameter k pada Imputasi

Salah satu temuan kunci dari proses optimasi GridSearchCV adalah nilai k optimal untuk KNNImputer yang bervariasi tergantung pada model klasifikasi yang digunakan. Hasil ini dirangkum dalam Tabel 3.5.

Tabel 2: Nilai k Optimal untuk KNN Imputer per Model

Model Klasifikasi	Nilai k Optimal untuk Imputasi	Model Klasifikasi
K-Nearest Neighbors (KNN)	11	K-Nearest Neighbors (KNN)
Support Vector Machine (SVM)	11	Support Vector Machine (SVM)
Random Forest	19	Random Forest

E. Analisis Keterbatasan: Ketidakseimbangan Kelas

Sebuah temuan yang konsisten dan signifikan di ketiga model adalah kegagalan total untuk mengidentifikasi kelas minoritas 'CL (Transplant)'. Seperti yang terlihat pada semua confusion matrix dan tabel metrik evaluasi (nilai presisi, recall, dan F1-score adalah 0.00), tidak ada

satu pun sampel dari kelas 'CL' di data uji yang berhasil diprediksi dengan benar.

Hal ini disebabkan oleh ketidakseimbangan kelas yang parah dalam dataset, di mana jumlah sampel kelas 'CL' (5 sampel pada data uji) sangat sedikit dibandingkan dengan kelas mayoritas 'C' (47 sampel) dan 'D' (32 sampel). Akibatnya, selama pelatihan, model belajar bahwa strategi paling efektif untuk memaksimalkan akurasi secara keseluruhan adalah dengan mengabaikan kelas minoritas. Keterbatasan ini menyoroti bahwa meskipun imputasi data berhasil dilakukan, performa prediktif sangat dipengaruhi oleh distribusi kelas pada data asli.

IV. KESIMPULAN & SARAN

A. Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang telah diuraikan pada bab sebelumnya, dapat ditarik tiga kesimpulan utama sebagai berikut:

i. Model Klasifikasi Terbaik:
Kombinasi metode K-Nearest Neighbor (KNN) Imputation dengan model klasifikasi Support Vector Machine (SVM) terbukti menjadi pendekatan yang paling efektif untuk memprediksi status kelangsungan hidup pasien sirosis pada dataset ini. Pendekatan ini berhasil mencapai akurasi tertinggi pada data uji sebesar 73.81% dan Weighted F1-Score sebesar 0.71, menunjukkan kinerja yang lebih unggul dan seimbang dibandingkan model KNN dan Random Forest.

ii. Pentingnya Optimasi Parameter Terintegrasi
Penelitian ini secara empiris menegaskan bahwa pemilihan parameter k dalam KNN Imputation adalah langkah kritis yang tidak dapat dipisahkan dari proses pemodelan. Nilai k optimal untuk imputasi terbukti bergantung pada konteks model klasifikasi yang digunakan setelahnya (misalnya, k=11 untuk SVM dan k=19 untuk Random Forest). Hal ini membuktikan bahwa pendekatan sistematis menggunakan GridSearchCV pada Pipeline terintegrasi lebih unggul daripada menggunakan heuristik statis, karena mampu menemukan sinergi terbaik antara tahap pra-pemrosesan dan klasifikasi.

iii. Keterbatasan Akibat Ketidakseimbangan Kelas
Kinerja prediktif dari semua model yang diuji secara signifikan dibatasi oleh adanya ketidakseimbangan kelas yang ekstrem dalam dataset. Hal ini menyebabkan kegagalan total dalam mengidentifikasi kelas minoritas ('CL' atau Transplant), di mana tidak ada satu pun sampel dari kelas tersebut yang berhasil diprediksi dengan benar. Keterbatasan ini menyoroti bahwa kualitas imputasi saja tidak cukup untuk menjamin kinerja model yang baik jika distribusi kelas data asli sangat timpang.

B. Saran untuk Penelitian Selanjutnya

Berdasarkan kesimpulan dan keterbatasan yang ditemukan, berikut adalah beberapa saran yang dapat menjadi acuan untuk penelitian selanjutnya di bidang ini:

i. Implementasi Penanganan Ketidakseimbangan Kelas
Untuk mengatasi masalah kegagalan prediksi pada kelas minoritas, sangat disarankan untuk menerapkan teknik resampling sebelum tahap pelatihan model. Metode seperti SMOTE (Synthetic Minority Over-sampling Technique) dapat digunakan untuk membuat sampel sintetis pada kelas minoritas, sehingga distribusi kelas menjadi lebih seimbang [22]. Mengintegrasikan SMOTE ke dalam pipeline penelitian diharapkan dapat meningkatkan nilai recall untuk kelas 'CL' secara signifikan.

ii. Studi Komparatif Metode Imputasi
Melakukan studi perbandingan yang lebih luas dengan mengevaluasi metode imputasi lain yang lebih canggih. Penelitian selanjutnya dapat membandingkan kinerja KNN Imputation dengan metode seperti MICE (*Multiple Imputation by Chained Equations*) [18] atau metode berbasis deep learning untuk menentukan pendekatan mana yang paling efektif dalam mempertahankan struktur data medis yang kompleks [19].

iii. Analisis Kepentingan Fitur (Feature Importance)
Untuk memberikan wawasan yang lebih bernilai secara klinis, disarankan untuk melakukan analisis kepentingan fitur menggunakan hasil dari model Random Forest. Analisis ini dapat mengidentifikasi variabel-variabel medis (misalnya, kadar Albumin, Bilirubin, atau Prothrombin) yang paling berpengaruh dalam memprediksi status kelangsungan hidup pasien, yang dapat berguna bagi para praktisi medis.

V. REFERENCES

[1] Dewi, K. P., et al. (2024). Hepatic Cirrhosis: A Literature Review. Jurnal Biologi Tropis, 24(1b), 35-41.
[2] Meidiansyah, R., & Herdayati, M. (2024). Risiko Kanker Hati pada Pasien dengan Penyakit Hati. Jurnal Kesehatan Masyarakat Mulawarman, 6(1), 13-23.
[3] Altamimi, A., et al. (2024). An automated approach to predict diabetic patients using KNN imputation. BMC Medical Research Methodology, 24, 221.
[4] Pradhan, O., et al. (2022). Lagged-kNN Based Data Imputation Approach for Multi-Stream Building Systems Data. International High Performance Buildings Conference.
[5] Sobarnas, M. A., & Iskandar. (2020). Komparasi Akurasi Metode K-Nearest Neighbor dan C4.5. Infotech: Jurnal Informatika & Teknologi, 1(1), 1-14.

[6] Prasetya, M. R. A., Priyatno, A. M., & Nurhaini, N. (2023). Penanganan Imputasi Missing Values pada Data Time Series. Jurnal Informasi dan Teknologi, 5(2), 56-62.
[7] Pamuji, F. Y., et al. (2024). Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil. Jurnal Informatika Polinema, 10(2), 257-264.
[8] Tampubolon, E., & Amin, R. (2024). Penerapan Algoritma K Nearest Neighbor untuk Prediksi Akurasi Penyakit Diabetes. Informatics and Digital Expert (INDEX).
[9] Zhang, L. (2020). A pattern-recognition-based ensemble data imputation framework for sensors. Sensors, 20(5974), 1-16.
[10] Amalia, M., et al. (2023). Karakteristik Pasien Sirosis Hepatis. UMI Medical Journal, 8(1), 53-61.
[11] Sania, A., et al. (2021). The K Nearest Neighbor Algorithm for Imputation of Missing Longitudinal Prenatal Alcohol Data. Columbia University Medical Center.
[12] Yusfila, F.Q., et al. (2025). Klasifikasi Penyakit Hepatitis C dengan Menggunakan K-Nearest Neighbor. Sains Data: Jurnal Studi Data dan Informasi Kesehatan.
[13] Nugroho, H., & Surendro, K. (2019). Missing Data Problem in Predictive Analytics. Proceedings of the 2019 7th International Conference on Software and Computer Applications, 95-100.
[14] Akakpo, S., et al. (2024). Optimization of the K-Nearest Neighbor Algorithm to Predict Bank Churn. Statistics, Optimization & Information Computing, 12, 1397-1408.
[15] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
[16] Akbar, A. S., & Kusumodestoni, R. H. (2020). Optimasi nilai k dan parameter lag algoritme k-nearest neighbor pada prediksi tingkat hunian hotel. Jurnal Teknologi dan Sistem Komputer, 8(3), 246-254.
[17] Arifianto, A. S., et al. (2022). PENGARUH PREDIKSI MISSING VALUE PADA KLASIFIKASI DECISION TREE C4.5. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 9(4), 779-786.
[18] Li, J., et al. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC medical research methodology, 24(1), 41.
[19] Liu, M., et al. (2021). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. arXiv preprint arXiv:2108.06734.
[20] Widyananda, W., et al. (2022). DATASET MISSING VALUE HANDLING AND CLASSIFICATION USING DECISION TREE C5.0 AND K-NN IMPUTATION: STUDY CASE CAR EVALUATION DATASET. Journal of Theoretical and Applied Information Technology, 100(12).
[21] Berrar, D. (2018). Cross-validation. In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (Vol. 1-3, pp. 542-545). Elsevier.
[22] Pamuji, F. Y., Dwi, S., & Putri, A. (2021). Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass. SENASIF, 331-338.