JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 2, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i2.9163

ANALISIS KOMPARATIF STRATEGI IMPUTASI NILAI HILANG PADA *DATASET* HEPATITIS UCI MENGGUNAKAN XGBOOST

Muhammad Mirza Kurniawan[1*], Betha Nurina Sari[2]

[1,2]Universitas Singaperbangsa Karawang; Jalan HS. Ronggo Waluyo, Telukjambe Timur, Karawang 41361, Jawa Barat, Indonesia; +62 267 641177

Keywords:
Imputasi Data Hilang;
XGBoost;
Hepatitis;
MICE;
KNN Imputation

Corespondent Email:
2210631170085@student.unsika.ac.id

Abstrak. Penyakit hepatitis masih menjadi tantangan kesehatan global yang signifikan, dengan beban kasus terbesar ditemukan di wilayah berkembang. Meskipun Rekam Kesehatan Elektronik (EHR) sangat bernilai bagi penelitian klinis dan pemodelan prediktif, data tersebut sering kali tidak lengkap. Laporan menunjukkan bahwa hingga 71% entri data dapat memiliki nilai hilang (*missing values*), yang menghadirkan tantangan substansial terhadap keandalan analisis data dan pembangunan model. Penelitian ini mengevaluasi efektivitas berbagai strategi imputasi data hilang pada *dataset* Hepatitis UCI, sebuah *benchmark* yang dikenal memiliki tingkat ketidaklengkapan tinggi. Kami membandingkan metode *Listwise deletion*, *Mean Imputation*, *K-Nearest Neighbors* (KNN), serta *Multivariate Imputation by Chained Equations* (*MICE*) beserta variannya. Evaluasi dilakukan menggunakan algoritma klasifikasi XGBoost dengan *Stratified 5-Fold Cross-Validation*. Hasil penelitian menunjukkan bahwa *Listwise deletion* tidak hanya mencapai kinerja rata-rata tertinggi dengan *F1-Score* sebesar 81,76%, tetapi juga menunjukkan stabilitas paling konsisten dengan standar deviasi terendah (6,22%) dibandingkan teknik imputasi kompleks lainnya yang menunjukkan variabilitas tinggi.

Abstract. *Hepatitis remains a significant global health challenge, with the greatest case burden found in developing regions. While Electronic Health Records (EHR) are highly valuable for clinical research and predictive modeling, such data are often incomplete. Reports indicate that up to 71% of data entries may contain missing values, posing substantial challenges to the reliability of data analysis and model development. This study evaluates the effectiveness of various missing data imputation strategies on the UCI Hepatitis dataset, a benchmark known for its high level of incompleteness. We compare Listwise deletion, Mean Imputation, K-Nearest Neighbors (KNN), and Multivariate Imputation by Chained Equations (MICE) along with their variants. Evaluation was conducted using the XGBoost classification algorithm with Stratified 5-Fold Cross-Validation. The results show that Listwise deletion not only achieved the highest average performance with an F1-Score of 81.76%, but also demonstrated the most consistent stability with the lowest standard deviation (6.22%) compared to other complex imputation techniques that exhibited high variability.*

1. PENDAHULUAN

Penyakit hepatitis tetap menjadi tantangan kesehatan global yang kritis, mempengaruhi sekitar 58 juta orang di seluruh dunia dengan hepatitis C dan berkontribusi terhadap morbiditas dan mortalitas yang

signifikan melalui sirosis hati dan karsinoma hepatoseluler [1]. Klasifikasi dini dan akurat dari tingkat keparahan penyakit hepatitis sangat penting untuk intervensi klinis yang tepat waktu dan peningkatan hasil pasien [2]. Pendekatan *machine learning* telah menunjukkan potensi yang cukup besar dalam mengembangkan model prediktif untuk klasifikasi hepatitis, memanfaatkan fitur klinis seperti usia, penanda biokimia, dan riwayat pasien untuk membedakan antara hasil penyakit yang berbeda [3], [4]. Namun, tantangan yang terus berlanjut dalam mengembangkan model prediktif yang kuat terletak pada penanganan nilai yang hilang yang sering terjadi dalam *dataset* klinis karena catatan pasien yang tidak lengkap, ketidaktersediaan tes laboratorium, dan kesalahan pengumpulan data [5].

Data yang hilang mewakili tantangan yang meluas dalam *dataset* kesehatan, dengan studi terbaru mendokumentasikan tingkat kehilangan yang sangat bervariasi di seluruh konteks klinis, dari sekitar 39,7% dalam beberapa *dataset* registri kanker hingga lebih dari 71,0% untuk jenis kanker tertentu dalam catatan kesehatan elektronik [6], [7]. Prevalensi nilai yang hilang menciptakan hambatan metodologis substansial yang dapat merusak validitas dan reliabilitas upaya pemodelan prediktif. Penanganan data yang hilang yang tidak memadai dapat mengakibatkan estimasi parameter yang bias, penurunan kekuatan statistik, dan pada akhirnya prediksi yang cacat yang dapat mempengaruhi proses pengambilan keputusan klinis secara merugikan [8], [9].

Terlepas dari kontribusi berbagai metode imputasi, kesenjangan penelitian yang kritis tetap ada. Penelitian sebelumnya [15], [16] terutama berfokus pada akurasi imputasi (MAE/RMSE), sering menggunakan *missingness* yang disimulasikan di bawah asumsi MCAR [16], menyimpulkan bahwa metode canggih seperti *MissForest* dan *MICE* mengungguli yang lebih sederhana. Namun, tidak jelas apakah keunggulan ini diterjemahkan ke kinerja klasifikasi akhir (*F1-Score*) ketika diterapkan pada *dataset* dengan pola kehilangan tinggi yang terjadi secara alami (MNAR). Selanjutnya, dalam skenario dengan tingkat kehilangan ekstrem (seperti *dataset* Hepatitis UCI, di mana 51,6% *record* tidak lengkap [17]), kelayakan strategi paling sederhana, *Listwise deletion*, sebagai *baseline* sering diabaikan daripada diuji secara ketat terhadap metode imputasi yang kompleks [10].

Oleh karena itu, penelitian ini bertujuan untuk secara ketat menantang asumsi bahwa imputasi kompleks selalu merupakan strategi superior dalam *dataset* klinis dengan tingkat kehilangan tinggi. Kami melakukan analisis komparatif komprehensif dari beberapa strategi penanganan pada *dataset* Hepatitis UCI, yang memiliki tingkat kehilangan parah yang terjadi secara alami (43,2% hilang di PROTIME) [17]. Tujuan penelitian spesifik adalah:

1. Untuk secara langsung membandingkan kinerja klasifikasi akhir (*F1-Score*) dari imputasi canggih yang di tuning secara ekstensif (*IterativeImputer*), imputasi sederhana (*Mean Imputation, KNN Imputation*), dan *baseline Listwise deletion* yang naif.
2. Untuk menentukan strategi mana yang secara faktual menghasilkan kinerja klasifikasi tertinggi menggunakan *classifier XGBoost* yang konstan dan 5-*Fold Cross-Validation* yang ketat.
3. Untuk menganalisis implikasi dari tingkat kehilangan tinggi (51,6% record tidak lengkap) pada prinsip *'Garbage In, Garbage Out'*, mempertanyakan apakah imputasi itu sendiri memperkenalkan *noise* berbahaya yang menurunkan kinerja lebih daripada hanya membuang data yang tidak lengkap.

Kontribusi utama dari penelitian ini adalah:

1. Perbandingan yang kuat dari imputasi multivariat yang di-*tuning* (*MICE*) terhadap *baseline Listwise deletion* yang sering diabaikan pada *dataset* dengan tingkat kehilangan tinggi.
2. Menyediakan bukti empiris bahwa untuk beberapa *dataset* dunia nyata yang rusak parah, *Listwise deletion* dapat secara signifikan mengungguli metode imputasi kompleks dalam tugas klasifikasi.

2. TINJAUAN PUSTAKA

2.1. Studi Terkait

Studi komparatif terbaru telah mengevaluasi berbagai metode imputasi untuk data kesehatan. Batra et al. [14] mengembangkan strategi *ensemble* menggunakan *Simple Mean, KNN*, dan *Iterative Imputation* untuk data COVID-19, mencapai kinerja yang lebih baik (misalnya, MAE

serendah 60,81 dalam satu kasus) daripada metode mandiri menggunakan data hilang dunia nyata. Li et al. [15] membandingkan delapan metode pada kohort penyakit kardiovaskular (CVD) (N = 10.164), menemukan bahwa KNN memiliki akurasi imputasi terbaik (MAE: 0,2032) sementara Random Forest (RF) menghasilkan kinerja prediktif tertinggi (AUC: 0,777). *MICE* berkinerja sedang (AUC: 0,720) dalam studi mereka, yang menggunakan data hilang yang disimulasikan. Joel et al. [16] menilai tujuh teknik di tiga *dataset* kesehatan, menyimpulkan bahwa *MissForest* secara konsisten berkinerja terbaik, diikuti oleh *MICE*, menggunakan data hilang yang disimulasikan. Sementara Li et al. [15] dan Joel et al. [16] terutama menggunakan data hilang yang disimu lasikan dan tidak secara khusus menguji IterativeImputer dari scikit-learn, Batra et al. [14] menggunakannya dalam ensemble mereka.

2.2. Strategi Penanganan Data Hilang

Berbagai metode imputasi telah diusulkan untuk menangani nilai yang hilang dalam data kesehatan.

2.2.1. Listwise deletion

Listwise deletion menghapus semua record yang tidak lengkap [10], menghasilkan *dataset* lengkap D_1^{complete} yang hanya terdiri dari baris-baris i di mana set fitur yang hilang x_i^{miss} adalah kosong $x_i^{miss} = \emptyset$. Keuntungannya meliputi kesederhanaan metodologis dan tidak adanya persyaratan pemodelan apa pun [10]. Namun, pendekatan ini secara substansial mengurangi ukuran sampel dan dapat memperkenalkan bias seleksi, terutama ketika data tidak *Missing Completely At Random* (MCAR) [10].

2.2.2. Mean Imputation

Mean imputation mengganti nilai numerik yang hilang dengan mean dari nilai yang diamati $\widetilde{x_{ij}}$ [10], [21], seperti yang ditunjukkan dalam persamaan (1); fitur kategorikal menggunakan modus:

$$\widetilde{x_{ij}} = \begin{cases} x_{ij} \text{jika } x_{ij} \text{ diamati} \\ \overline{x_j} = \frac{1}{|O_j|}\sum_{k\in O_j} x_{kj} \text{ jika } x_{ij} \text{ hilang} \end{cases} \quad (1)$$

2.2.3. K-Nearest Neighbors (KNN) Imputation

KNN mengimputasi nilai yang hilang menggunakan kesamaan lokal [22]. Pertama, ia menghitung jarak *Euclidean* antara *record* hanya menggunakan fitur yang diamati, seperti yang dijelaskan dalam persamaan (2):

$$d(x_i, x_m) = \sqrt{\sum_{j'\in O_i \cap O_m} \left(x_{ij'} - x_{mj'}\right)^2} \quad (2)$$

Kemudian, ia mengimputasi nilai yang hilang $\widetilde{x_{ij}}$ sebagai rata-rata tertimbang dari k tetangga terdekatnya, dihitung menggunakan persamaan (3):

$$\widetilde{x_{ij}} = \frac{\sum_{m\in N_k(x_i)} w_m \cdot x_{mj}}{\sum_{m\in N_k(x_i)} w_m}, \quad w_m = \frac{1}{d(x_i, x_m)+\varepsilon} \quad (3)$$

Metode ini mempertahankan struktur lokal tetapi sensitif terhadap k dan tingkat kehilangan yang tinggi [23].

2.2.4. Imputasi Lanjutan (MICE)

Metode imputasi berbasis *machine learning* yang lebih canggih telah muncul untuk mengatasi keterbatasan ini. *Multiple Imputation by Chained Equations* (*MICE*), di implementasikan sebagai *IterativeImputer* di *scikit-learn*, melakukan imputasi multivariat dengan memodelkan setiap fitur dengan nilai yang hilang secara iteratif sebagai fungsi dari fitur lainnya [12], [13].

2.3. XGBoost Classifier

Algoritma XGBoost dikenal memiliki keunggulan dalam hal kecepatan, skalabilitas, efisiensi, dan kesederhanaan dalam pemodelan prediktif [26]. Prediksi $\widehat{y_i}$ untuk sampel x_i adalah jumlah prediksi dari K pohon, seperti yang diberikan oleh persamaan (4) [24]:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) \quad (4)$$

Setiap pohon f_t dilatih secara berurutan untuk meminimalkan fungsi tujuan yang diatur $L^{(t)}$, yang mencakup istilah kerugian l dan istilah regularisasi $\Omega(f_t)$, didefinisikan dalam persamaan (5):

$$L^{(t)} = \sum_{i=1}^{N} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t), \quad \Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \quad (5)$$

XGBoost dipilih karena efektivitasnya yang terbukti pada data tabular medis untuk tugas klasifikasi [25].

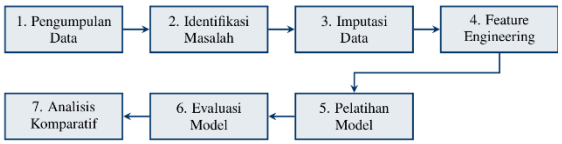
3. METODE PENELITIAN

3.1. Kerangka Penelitian

Penelitian ini menggunakan kerangka analisis komparatif sistematis untuk mengevaluasi efektivitas beberapa strategi

penanganan nilai hilang yang berbeda pada klasifikasi penyakit hepatitis menggunakan *dataset* Hepatitis UCI. Penelitian ini mengikuti *pipeline* eksperimental terstruktur, yang digambarkan dalam Gambar 1, memastikan penilaian yang adil dan konsisten di semua metode. Alur kerja mencakup tujuh tahap utama:

1. Tinjauan pustaka untuk mengidentifikasi teknik imputasi mutakhir dan menetapkan kesenjangan penelitian [14], [15], [16];
2. Akuisisi data dan deskripsi *dataset* Hepatitis UCI;
3. Pra-pemrosesan data yang melibatkan pengkodean dan penskalaan;
4. Implementasi dan eksekusi lebih dari sepuluh pipeline imputasi/penanganan paralel, ter masuk tiga *baseline* umum (*Listwise deletion, Mean Imputation, KNN Impu0074ation*) dan tuning hyperparameter ekstensif pada beberapa varian metode IterativeImputer (*MICE*);
5. Pelatihan model klasifikasi XGBoost yang konstan pada output dari setiap pipeline;
6. Evaluasi kinerja menggunakan *Stratified* 5-*Fold Cross-Validation* dan metrik klasifikasi standar;
7. Analisis komparatif untuk mengidentifikasi strategi imputasi yang superior berdasarkan kinerja klasifikasi, khususnya *F1-Score*.



Gambar 1. Kerangka Penelitian

Inti dari metodologi ini terletak pada isolasi dampak teknik penanganan nilai yang hilang. Dengan menjaga *dataset*, langkah-langkah pra-pemrosesan (di luar imputasi), model klasifikasi (XGBoost), dan protokol evaluasi (*Stratified* K-*Fold* CV, metrik) konstan di keempat pipeline, setiap perbedaan yang diamati dalam kinerja klasifikasi dapat diatribusikan secara langsung pada efektivitas strategi nilai hilang masing-masing [15], [18], [19].

3.2. Deskripsi Dataset

Penelitian ini menggunakan *dataset* Hepatitis yang tersedia untuk umum dari UCI Machine Learning Repository, yang awalnya dikumpulkan oleh G. Gong [17]. *Dataset* ini terdiri dari 155 catatan pasien, masing-masing dijelaskan oleh 20 atribut klinis, yang bertujuan untuk memprediksi hasil kelangsungan hidup biner (Kelas: 1=DIE, 2=LIVE).

Karakteristik utama dari *dataset* ini, yang membuat nya cocok untuk studi imputasi ini, adalah adanya nilai hilang yang terjadi secara alami di berbagai atribut, meniru tantangan data klinis dunia nyata [5]. Rincian atribut *dataset* dan distribusi nilai hilang disajikan dalam Tabel 1. Tingkat kehilangan yang signifikan diamati pada PROTIME (67 hilang, 43,23%), ALBUMIN (16 hilang, 10,32%), dan ALK PHOSPHATE (29 hilang, 18,71%), antara lain. Akibatnya, hanya 75 record (48,39%) yang lengkap, sementara 80 record (51,61%) menunjukkan setidaknya satu entri yang hilang. Distribusi variabel target menunjukkan ketidakseimbangan kelas moderat, dengan 32 contoh kelas ‘DIE’ (20,65%) dan 123 contoh kelas ‘LIVE’ (79,35%). Ketidakseimbangan ini mengharuskan penggunaan pengambilan sampel bertingkat (*Stratified* sampling) selama evaluasi dan memprioritaskan metrik *F1-Score*.

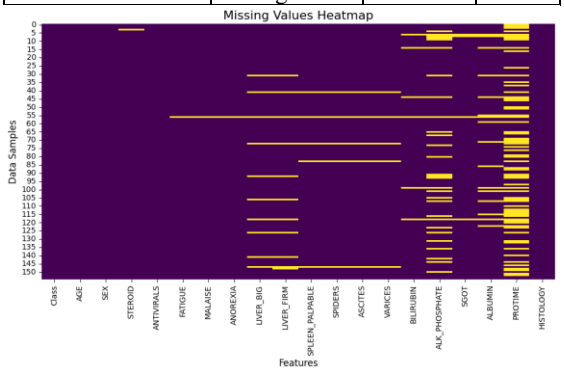
3.2.1. Analisis Distribusi Nilai Hilang

Langkah pertama dalam penelitian ini adalah melakukan analisis mendalam terhadap distribusi nilai hilang dalam *dataset*. Gambar 2 menyajikan visualisasi heatmap yang menunjukkan pola nilai hilang di seluruh fitur dan sampel data.

Tabel 1. Distribusi Nilai Hilang pada *Dataset* Hepatitis UCI

Atribut	Tipe	Hilang	(%)
PROTIME	Kontinu	67	43.23
ALK_PHOSPHATE	Kontinu	29	18.71
ALBUMIN	Kontinu	16	10.32
LIVER_FIRM	Kategorikal	11	7.10
LIVER_BIG	Kategorikal	10	6.45
BILIRUBIN	Kontinu	6	3.87
SPLEEN_PALPABLE	Kategorikal	5	3.23
VARICES	Kategorikal	5	3.23
ASCITES	Kategorikal	5	3.23
SPIDERS	Kategorikal	5	3.23
SGOT	Kontinu	4	2.58

Atribut	Tipe	Hilang	(%)
ANOREXIA	Kategorikal	1	0.65
STEROID	Kategorikal	1	0.65
FATIGUE	Kategorikal	1	0.65
MALAISE	Kategorikal	1	0.65



Gambar 2. *Heatmap* Nilai Hilang pada *Dataset* Hepatitis UCI

Heatmap secara visual mengonfirmasi keberadaan nilai hilang yang tersebar tidak merata di berbagai fitur. Intensitas warna yang berbeda dengan jelas menunjukkan bahwa beberapa fitur memiliki kepadatan entri yang hilang jauh lebih tinggi daripada yang lain. Distribusi yang tidak seragam ini memiliki implikasi penting untuk pemilihan strategi imputasi.

3.3. Identifikasi Masalah

Misalkan $D = \{(x_i, y_i)\}_{i=1}^{N}$ mewakili *dataset* hepatitis, di mana $x_i \in R^d$, $y_i \in \{0,1\}$, $N = 155$, $d = 20$. Karena nilai yang hilang, data yang diamati adalah $D_{obs} = \{(x_i^{obs}, x_i^{miss}, y_i)\}_{i=1}^{N}$. Kami membandingkan empat strategi imputasi $I = \{I_1, I_2, I_3, I_4\}$, di mana setiap strategi I_j mentransformasikan D_{obs} menjadi *dataset* lengkap $D_j^{complete} = \{(\widetilde{x_i^j}, y_i)\}_{i=1}^{N_j}$. Di sini, $\widetilde{x_i^j}$ Adalah vektor yang diimputasi dan N_j adalah ukuran sampel yang dipertahankan ($N_1 < N$ untuk *listwise deletion*, $N_2 = N_3 = N_4 = N$ untuk lainnya). Tujuannya adalah untuk mengidentifikasi metode imputasi terbaik I_j^* yang didefinisikan secara formal sebagai:

$$I_j^* = \arg\max_{I_j \in I} \text{Performance}\left(f\left(D_j^{complete}\right)\right) \quad (6)$$

di mana *performance* dalam persamaan (6) diukur menggunakan akurasi, presisi, *recall*, dan F1-Score melalui *Stratified* 5-*Fold Cross-Validation* [20].

3.4. Skenario Eksperimen

Selain tiga metode *baseline* yang didefinisikan sebelumnya, penelitian ini juga mengevaluasi *Iterative Imputer* (*MICE*). Berbeda dengan *baseline*, yang memiliki konfigurasi tetap, kinerja *MICE* sangat bergantung pada *hyperparameter* internal dan estimatornya.

Untuk memastikan perbandingan yang kuat dan mengidentifikasi potensi maksimum dari imputasi canggih, penelitian ini tidak mengandalkan pengaturan default. Kami melakukan eksperimen *tuning hyperparameter* yang ekstensif untuk menemukan konfigurasi *MICE* yang optimal untuk *dataset* spesifik ini. Eksperimen ini melibatkan pengujian lebih dari sepuluh varian *MICE* yang berbeda, termasuk:

1. Menyesuaikan parameter dasar (misalnya, max_iter, initial_strategy, imputation_order).
2. Menguji estimator internal yang berbeda (misalnya, *Bayesian Ridge*, *Extra Trees Regressor*, *Random Forest Regressor*) untuk menangani hubungan non-linear, meniru pendekatan *MissForest* yang ditemukan efektif dalam studi lain [16].
3. Mengembangkan strategi *hybrid* baru (misalnya, *Hybrid MICE + KNN*, *Hybrid MICE + Mean*, dan *MICE + Missing Indicators*).

Implementasi spesifik dari strategi yang digunakan dalam eksperimen ini adalah:

- *Listwise deletion*: Menghapus baris dengan nilai hilang (*Baseline*).
- Mean Imputation: Menggunakan rata-rata untuk numerik dan modus untuk kategorikal.
- KNN Imputation: Menggunakan k = 5 berdasarkan bukti empiris dari studi sebelumnya [14], [15].
- XGBoost Hyperparameters: Konfigurasi *hyperparameter* model disajikan pada Tabel 2.

Tabel 2. Hasil Kinerja Metode Imputasi dengan *Classifier*

Hyperparameter	**Nilai**	**Deskripsi**
n_estimators	100	Jumlah pohon keputusan yang dibangun.

Hyperparameter	Nilai	Deskripsi
max_depth	5	Kedalaman maksimum setiap pohon.
learning_rate	0.1	Kontribusi setiap pohon (*step size*).
subsample	0.8	Proporsi sampel data untuk setiap pohon.
colsample_bytree	0.8	Proporsi fitur untuk setiap pohon.

Kinerja klasifikasi dari semua varian ini selanjutnya dievaluasi terhadap tiga *baseline* (*Listwise deletion, Mean Imputation, KNN Imputation*) untuk secara faktual menentukan strategi berkinerja tertinggi.

3.5. Strategi Evaluasi

3.5.1. Stratified 5-Fold Cross-Validation

Dataset D dipartisi menjadi $K = 5$ *Fold*, D_k. Untuk setiap *Fold* k, set pelatihan adalah $D_{train}^{(k)} = D \setminus D_k$ dan set uji adalah $D_{test}^{(k)} = D_k$. Metrik kinerja rata-rata M_j di seluruh *Fold* dihitung menggunakan persamaan (7):

$$M_j = \frac{1}{K}\sum_{k=1}^{K} M_j^{(k)} \quad (7)$$

Memastikan distribusi kelas asli (20,6% DIE, 79,4% LIVE) dipertahankan di setiap *Fold* [17].

3.5.2. Metrik Kinerja

Empat metrik klasifikasi standar yang diturunkan dari matriks kebingungan (TP, TN, FP, FN) digunakan: Akurasi, seperti yang didefinisikan dalam persamaan (8):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Presisi, seperti yang didefinisikan dalam persamaan (9):

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

Recall, seperti yang didefinisikan dalam persamaan (10):

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

F1-Score, seperti yang didefinisikan dalam persamaan (11):

$$F1\text{-}Score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (11)$$

F1-Score adalah metrik utama untuk membandingkan secara adil dan akurat efektivitas setiap metode imputasi nilai hilang pada tugas klasifikasi akhir [25].

4. HASIL DAN PEMBAHASAN

4.1. Hasil Kinerja Komparatif

Tabel 3 menyajikan hasil kinerja komprehensif dari sepuluh *pipeline* eksperimental yang diuji dalam penelitian ini. Setiap pipeline mewakili kombinasi strategi imputasi dan classifier XGBoost, dievaluasi menggunakan *Stratified* 5 *Fold Cross-Validation*. Hasil menunjukkan bahwa *Listwise deletion* mencapai kinerja tertinggi dengan akurasi 83.59%, presisi 83.89%, *recall* 82.78%, dan *F1-Score* 81.76%. Hal ini merupakan temuan yang berlawanan dengan intuisi umum yang mengasumsikan bahwa metode imputasi canggih akan menghasilkan kinerja lebih baik.

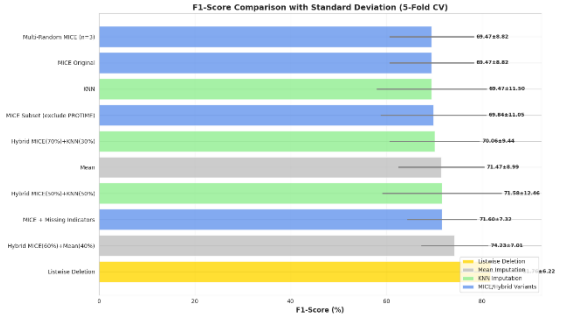
Tabel 3. Hasil Kinerja Metode Imputasi dengan *Classifier* XGBoost

Metode Imputasi	Acc (%)	Prec (%)	Rec (%)	F1 (%)
Listwise deletion	83.59	83.89	82.78	**81.76**
Hybrid MICE(60%)+Mean(40%)	79.96	73.09	77.21	74.25
MICE + Missing Indicators	77.95	68.83	79.91	72.53
Hybrid MICE(50%)+KNN(50%)	77.32	68.99	78.14	72.01
Mean Imputation	76.73	68.70	74.94	70.87
Hybrid MICE(70%)+KNN(30%)	76.07	67.96	75.56	70.44
MICE Subset (exclude PROTIME)	76.04	65.44	79.00	70.41
KNN Imputation (k=5)	75.45	65.98	76.56	69.61
MICE Original	74.82	67.08	72.40	68.68
Multi-Random *MICE* (n=3)	74.20	65.51	74.99	68.58

4.2. Visualisasi Perbandingan Kinerja

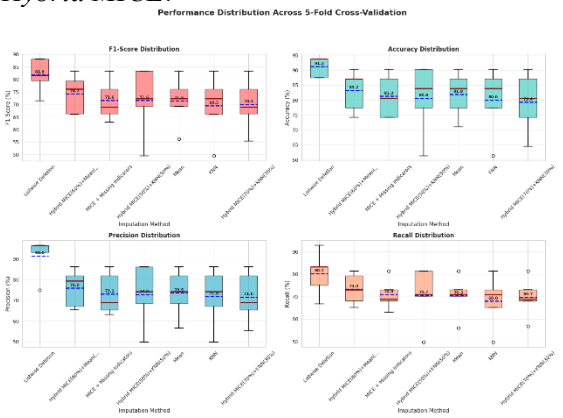
Gambar 3 menyajikan perbandingan terperinci *F1-Score* di semua metode imputasi, dilengkapi dengan error bar yang menunjukkan standar deviasi. Grafik ini menegaskan

dominasi *Listwise deletion* tidak hanya dalam rata-rata kinerja tetapi juga stabilitas relatifnya, di mana metode ini menunjukkan kinerja rata-rata tertinggi (81.76%) dengan standar deviasi yang terkendali.



Gambar 3. Perbandingan *F1-Score* (%) dengan Standar Deviasi.

Untuk menganalisis sebaran kinerja di seluruh *Fold* validasi silang, kami menyajikan distribusi kinerja pada Gambar 4. Visualisasi ini menunjukkan variabilitas *Listwise deletion* (rentang 71%-88%), yang mengindikasikan bahwa meskipun rata-ratanya tinggi, metode ini sensitif terhadap subset data uji tertentu, namun tetap lebih unggul dibandingkan variabilitas *Hybrid MICE*.



Gambar 4. Distribusi Kinerja (*Boxplot*) di 5-*Fold Cross-Validation*.

Untuk memberikan transparansi penuh terhadap variabilitas yang divisualisasikan di atas, Tabel 4 menyajikan rincian skor F1 untuk setiap *Fold* validasi. Data ini menegaskan bahwa meskipun *Listwise deletion* mengalami penurunan kinerja pada *Fold* 5 (71.43%), ia tetap mengungguli metode *Hybrid* yang jatuh lebih dalam (hingga 49.46%) pada kondisi yang sama

Tabel 4. Rincian *F1-Score* (%) per *Fold* Eksperimen untuk 5 Metode Teratas

Fold	**Listwise**	***Hybrid MICE***	***MICE* +Ind**	***Hybrid* KNN**	**Mean Imp**
Fold 1	81.61	83.42	83.42	83.42	83.42
Fold 2	79.49	79.33	69.00	72.37	72.37
Fold 3	**88.15**	76.15	76.15	83.42	76.15
Fold 4	**88.15**	66.30	66.30	69.22	56.20
Fold 5	71.43	65.93	63.10	49.46	69.22
Mean	**81.76**	74.23	71.60	71.58	71.47
Std Dev	**6.22**	7.01	7.32	12.46	8.99

4.3. Analisis Statistik dan Variabilitas

Analisis mendalam terhadap variabilitas hasil pada Gambar 4 mengungkapkan wawasan penting yang sering terlewatkan jika hanya melihat nilai rata-rata. Observasi ini terbagi menjadi dua aspek utama:

1. **Stabilitas *Listwise deletion***
   Meskipun menggunakan sampel yang jauh lebih sedikit (N=75), metode *Listwise deletion* mempertahankan standar deviasi sebesar **6.22%**. Nilai ini lebih rendah dibandingkan *Hybrid MICE(60%)+Mean(40%)* (7.01%) dan *Hybrid KNN* (12.46%). Temuan ini membantah anggapan umum bahwa pengurangan ukuran sampel secara drastis akan selalu meningkatkan varians model pada *dataset* ini.
2. **Ketidakstabilan Metode Imputasi**
   Metode *Hybrid MICE(50%)+KNN(50%)*, meskipun merupakan strategi yang canggih, menunjukkan rentang kinerja yang sangat lebar (Min 49.46% – Max 83.42%) dengan *Coefficient of Variation* (CV) tertinggi sebesar **17.41%**. Hal ini mengindikasikan bahwa metode imputasi yang kompleks dapat memperkenalkan ketidakpastian yang signifikan tergantung pada lipatan data (*Fold*) yang digunakan.

4.4. Analisis Temuan Kunci

Analisis menyeluruh dari hasil eksperimental mengungkapkan dua temuan kunci utama:

1. **Superioritas *Listwise deletion***
   Metode ini secara konsisten mengungguli semua metode imputasi lainnya, mencapai *F1-Score* sebesar **81.76%**. Sebagai perbandingan,

pipeline terbaik berikutnya yaitu *Hybrid MICE(60%)+Mean(40%)* hanya mencapai 74.25%.

2. **Efektivitas Metode *Hybrid***
   Metode *hybrid* yang menggabungkan *MICE* dengan teknik lain (Mean atau KNN) cenderung berkinerja lebih baik daripada implementasi *MICE* murni.

4.5. Paradoks Listwise deletion

Hasil penelitian ini menghadirkan paradoks yang berlawanan dengan intuisi: pada *dataset* Hepatitis UCI, strategi penanganan nilai hilang yang paling sederhana (*Listwise deletion*) justru mengungguli teknik imputasi yang jauh lebih canggih. Temuan ini bertentangan dengan rekomendasi umum dalam literatur yang menyarankan penggunaan metode imputasi untuk memaksimalkan retensi informasi.

4.5.1. Implikasi Tingginya Tingkat Data Hilang

Dataset Hepatitis ditandai dengan tingkat data hilang (*missingness*) yang sangat tinggi, dengan beberapa atribut mencapai **43.2%** ketidaklengkapan. Dalam skenario seperti ini, metode imputasi dipaksa untuk mensintesis proporsi besar dari setiap fitur. Estimasi yang diimputasi ini bukanlah data aktual, melainkan prediksi yang memperkenalkan variasi artifisial dan potensi bias ke dalam *dataset*. Temuan ini sejalan dengan studi klasifikasi medis sebelumnya di mana tingkat *missingness* secara signifikan memengaruhi kinerja model [6], [8]. Dalam kasus *dataset* Hepatitis, tingkat ketidaklengkapan yang ekstrem menyebabkan degradasi kualitas imputasi.

4.5.2. Prinsip "Garbage In, Garbage Out"

Ketika lebih dari sepertiga nilai sebuah fitur tidak ada, urutan kausal menjadi terbalik. Alih-alih menggunakan data nyata untuk memprediksi hasil, model menggunakan nilai yang diprediksi untuk memprediksi hasil akhir. Hal ini mereduksi model menjadi rantai asumsi yang sangat bergantung pada premis awal yang benar. Lin et al. [19] membahas bahwa metode imputasi *deep learning* untuk data kontinu memiliki asumsi tentang diskretisasi data. Ketika asumsi ini dilanggar, yang sangat mungkin terjadi pada *dataset* klinis dunia nyata, kinerja imputasi dapat menurun drastis.

4.5.3. Kelebihan Listwise deletion pada Konteks Ini

Dengan menghapus *record* yang tidak lengkap, *Listwise deletion* memastikan bahwa *classifier* dilatih secara eksklusif pada observasi yang benar-benar diamati (data riil). Meskipun langkah ini secara substansial mengurangi ukuran sampel (dari 155 menjadi sekitar 80 *record*), data yang tersisa memiliki integritas tinggi. Hal ini kontras dengan *dataset* hasil imputasi yang memiliki ukuran lebih besar tetapi mengandung proporsionalitas *noise* sintetis yang lebih tinggi.

4.6. Perbandingan dengan Penelitian Terkait

Hasil penelitian ini konsisten dengan beberapa studi yang menunjukkan bahwa metode imputasi sederhana terkadang dapat mengungguli metode kompleks pada *dataset* tertentu [15], [16]. Metrik yang digunakan dalam evaluasi ini mengikuti praktik standar yang telah digunakan dalam penelitian sebelumnya [20], [25]. Namun, hasil studi ini memiliki karakteristik unik karena adanya kombinasi tiga faktor: ukuran sampel yang kecil, tingkat ketidaklengkapan yang tinggi (hingga 43.2%), dan ketidakseimbangan kelas pada *dataset* Hepatitis.

4.7. Implikasi Praktis

Temuan ini memiliki implikasi penting untuk praktik penanganan data klinis:

1. **Ambang Batas Penggunaan**
   Untuk *dataset* dengan tingkat ketidaklengkapan sangat tinggi (>30%), *Listwise deletion* mungkin menjadi pilihan yang lebih aman dan robust dibandingkan imputasi kompleks.
2. **Pentingnya Evaluasi Empiris**
   Evaluasi komparatif terhadap berbagai strategi imputasi harus dilakukan pada setiap *dataset* spesifik sebelum menentukan metode akhir, tidak bisa hanya mengandalkan asumsi teoritis.
3. **Kualitas vs Kuantitas**
   Kualitas data yang sebenarnya diamati (*observed data*) terbukti lebih penting daripada sekadar mengejar kuantitas data melalui proses imputasi.

5. KESIMPULAN

Penelitian ini menyajikan evaluasi empiris komprehensif mengenai efektivitas berbagai

strategi imputasi nilai hilang pada *dataset* Hepatitis UCI dengan menggunakan algoritma klasifikasi XGBoost. Berdasarkan hasil eksperimen dan pembahasan yang telah dilakukan, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. **Superioritas *Listwise deletion***
   Metode *Listwise deletion* terbukti menjadi pendekatan yang paling efektif, mencapai kinerja tertinggi dengan *F1-Score* sebesar **81.76%**. Metode ini mengungguli seluruh metode imputasi lainnya yang diuji dengan selisih kinerja yang signifikan, minimal sebesar 7.51%.
2. **Inefektivitas Imputasi pada Missingness Tinggi**
   Penerapan metode imputasi canggih seperti *MICE* dan KNN tidak menjamin peningkatan kinerja, terutama pada *dataset* dengan tingkat ketidaklengkapan yang sangat tinggi (>40% pada beberapa atribut).
3. **Kinerja Metode *Hybrid*:**Metode *hybrid* yang menggabungkan *MICE* dengan teknik sederhana (seperti *Mean* atau KNN) cenderung menghasilkan kinerja yang lebih baik dibandingkan dengan implementasi *MICE* murni.
4. **Prioritas Kualitas Data:** Hasil penelitian menegaskan bahwa kualitas data yang benar-benar diamati (*observed data*) memegang peranan yang lebih krusial terhadap kinerja klasifikasi dibandingkan dengan kuantitas data yang diperoleh melalui proses imputasi.

5.1. Kontribusi Penelitian

Kelebihan dan kontribusi utama dari penelitian ini meliputi: (1) evaluasi komprehensif terhadap sepuluh strategi imputasi yang berbeda, (2) penggunaan validasi silang (*Cross-Validation*) untuk memastikan hasil yang *robust*, dan (3) identifikasi paradoks penting dalam penanganan data hilang di mana metode sederhana dapat mengungguli metode kompleks pada kondisi tertentu.

5.2. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan:

- Fokus evaluasi hanya terbatas pada satu *dataset* (Hepatitis UCI).
- Pengujian hanya menggunakan satu jenis *classifier* (XGBoost).
- Belum dilakukannya eksplorasi mendalam mengenai *threshold* (ambang batas) optimal tingkat data hilang.

5.3. Saran Pengembangan Selanjutnya

Berdasarkan temuan dan keterbatasan tersebut, arah penelitian masa depan disarankan untuk:

1. Mengeksplorasi *threshold* optimal persentase data hilang di mana teknik imputasi mulai menjadi kontraproduktif dibandingkan penghapusan data.
2. Menginvestigasi efektivitas metode imputasi berbasis *Deep Learning* (seperti *GAIN* atau *VAE*).
3. Melakukan validasi metode pada *dataset* klinis lain dengan karakteristik berbeda untuk menguji generalisasi temuan ini.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] Al-Amain, F. T. Janin, F. Ahmed Robin, S. Ahmed, and K. M. Mohi Uddin, "Unleashing Machine Learning for Hepatitis C Prediction: A Holistic Exploration of Clinical Insights," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, 2024, pp. 1–6. doi: 10.1109/COMPAS60761.2024.10796087.

[2] Y. Fan, X. Lu, and G. Sun, "IHCP: interpretable hepatitis C prediction system based on black-box machine learning models," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–16, 2023, doi: 10.1186/s12859-023-05456-0.

[3] A. Alizargar, Y. L. Chang, and T. H. Tan, "Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques," *Bioengineering*, vol. 10, no. 4, 2023, doi: 10.3390/bioengineering10040481.

[4] M. O. Edeh *et al.*, "Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease," *Front. Public Heal.*, vol. 10, no. April, 2022, doi: 10.3389/fpubh.2022.892371.

[5] A. M. Elsayad, A. M. Nassef, and M. Al-Dhaifallah, "Diagnosis of Hepatitis Disease with Logistic Regression and Artificial Neural Networks," *J. Comput. Sci.*, vol. 16, no. 3, pp. 364–377, Mar. 2020, doi: 10.3844/jcssp.2020.364.377.

[6] D. X. Yang *et al.*, "Prevalence of Missing Data in the National Cancer Database and Association with Overall Survival," *JAMA Netw. Open*, vol. 4, no. 3, 2021, doi: 10.1001/jamanetworkopen.2021.1793.

[7] N. Cesare and L. P. O. Were, "A multi-step approach to managing missing data in time and patient variant electronic health records," *BMC Res. Notes*, vol. 15, no. 1, pp. 1–7, 2022, doi: 10.1186/s13104-022-05911-w.

[8] B. Bouvarel, F. Carrat, and N. Lapidus, "Updating mortality risk estimation in intensive care units from high-dimensional electronic health records with incomplete data," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, pp. 1–9, 2023, doi: 10.1186/s12911-023-02264-7.

[9] J. Mi, R. D. Tendulkar, S. M. C. Sittenfeld, S. Patil, and E. C. Zabor, "Combining Missing Data Imputation and Internal Validation in Clinical Risk Prediction Models," *Stat. Med.*, vol. 44, no. 18–19, pp. 1–15, 2025, doi: 10.1002/sim.70203.

[10] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, *A survey on missing data in machine learning*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00516-9.

[11] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured *datasets*: a systematic review," *BMC Med. Res. Methodol.*, vol. 24, no. 1, 2024, doi: 10.1186/s12874-024-02310-6.

[12] I. El Badisy, N. Graffeo, M. Khalis, and R. Giorgi, "Multi-metric comparison of machine learning imputation methods with application to breast cancer survival," *BMC Med. Res. Methodol.*, vol. 24, no. 1, 2024, doi: 10.1186/s12874-024-02305-3.

[13] Z. Chen, S. Tan, U. Chajewska, C. Rudin, and R. Caruana, "Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help?," *Proc. Mach. Learn. Res.*, vol. 209, pp. 86–99, 2023.

[14] S. Batra, R. Khurana, M. Z. Khan, W. Boulila, A. Koubaa, and P. Srivastava, "A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records," *Entropy*, vol. 24, no. 4, pp. 1–20, 2022, doi: 10.3390/e24040533.

[15] J. H. Li *et al.*, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study *datasets*," *BMC Med. Res. Methodol.*, vol. 24, no. 1, pp. 1–9, 2024, doi: 10.1186/s12874-024-02173-x.

[16] L. O. Joel, W. Doorsamy, and B. S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic *datasets*," *Int. J. Data Sci. Anal.*, vol. 20, no. 7, pp. 6357–6373, 2025, doi: 10.1007/s41060-025-00825-9.

[17] UCI Machine Learning Repository, "Hepatitis." 1983. [Online]. Available: https://doi.org/10.24432/C5Q59J

[18] L. Jin *et al.*, "A comparative study of evaluating missing value imputation methods in label-free proteomics.," *Sci. Rep.*, vol. 11, no. 1, p. 1760, Jan. 2021, doi: 10.1038/s41598-021-81279-4.

[19] W.-C. Lin, C.-F. Tsai, and J. R. Zhong, "Deep learning for missing value imputation of continuous data and the effect of data discretization," *Knowledge-Based Syst.*, vol. 239, p. 108079, 2022, doi: https://doi.org/10.1016/j.knosys.2021.108079.

[20] S. Wu, W. Yau, T. Ong, and S.-C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for

Telco Business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusId:233434157

[21] H. Rosado-Galindo and S. Dávila-Padilla, "Tree-Based Missing Value Imputation Using Feature Selection," *J. Data Sci.*, vol. 18, no. 4, pp. 606–631, 2020, doi: 10.6339/JDS.202010_18(4).0002.

[22] K. M. Fouad, M. M. Ismail, A. T. Azar, and M. M. Arafa, "Advanced methods for missing values imputation based on similarity learning.," *PeerJ. Comput. Sci.*, vol. 7, p. e619, 2021, doi: 10.7717/peerj-cs.619.

[23] M. S. Santos, P. H. Abreu, S. Wilk, and J. Santos, "How distance metrics influence missing data imputation with k-nearest neighbours," *Pattern Recognit. Lett.*, vol. 136, pp. 111–119, 2020, doi: 10.1016/j.patrec.2020.05.032.

[24] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[25] D. Rohmayani, C. A. Sugianto, R. S. Perdana, and M. Mansoor, "Improving Extreme Gradient Boosting Model for Heart Disease Prediction Using SMOTE for Class Imbalance," vol. 6, no. 4, pp. 1717–1728, 2025.

[26] F. H. Syahadah, R. T. Subagio, and P. Rizqiyah, "Penerapan XGBoost dalam Prediksi Pendaftaran Siswa Baru Bimbingan Belajar QSC di Kota Cirebon," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 3S1, pp. 1082–1089, 2025.