Journal of Artificial Intelligence and Software Engineering
Vol. 6, No. 2, Juni 2026, pp. 118~129
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v6i2.9409

Information Gain and Random Forest for Sex Classification Based on Craniometric Measurements

Nabilla Alya Firana[1], Iis Afrianty[2*], Novriyanto[3], Febi Yanto[4]
[1,2,3,4] Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia

Informasi Artikel

Diterima : 15 Juni 2026
Revisi : 22 Juni 2026
Publikasi : 30 juni 2026

Kata Kunci:

Antropologi Forensik
Information Gain
Klasifikasi Jenis Kelamin
Kraniometrik
Random Forest

ABSTRAK

Identifikasi jenis kelamin dari tulang tengkorak merupakan aspek penting dalam antropologi forensik, namun metode tradisional masih menghadapi keterbatasan berupa subjektivitas penilaian dan variasi antar populasi. Penelitian ini mengusulkan penerapan *Information Gain* sebagai teknik seleksi fitur dan *Random Forest* sebagai algoritma klasifikasi untuk menentukan jenis kelamin berdasarkan data kraniometrik. Dataset yang digunakan adalah dataset Howells yang terdiri dari 2.524 sampel dengan 83 fitur pengukuran tulang tengkorak. Proses seleksi fitur menggunakan *Information Gain* dilakukan dengan variasi nilai *threshold* 0,01; 0,05; dan 0,09, dilanjutkan pengujian tambahan pada rentang *threshold* 0,01 hingga 0,09. Evaluasi model menggunakan metode 10-Fold *Cross Validation* dengan parameter default algoritma *Random Forest*. Hasil pengujian menunjukkan bahwa *threshold* 0,02 menghasilkan 57 fitur terpilih dari 83 fitur awal dengan performa terbaik, yaitu *accuracy* sebesar 87,40%, *precision* 87,53%, *recall* 87,40%, dan *F1-score* 87,41%. Hasil ini meningkat dibandingkan model *baseline* tanpa seleksi fitur yang menghasilkan *accuracy* 86,57%. Penelitian ini menunjukkan bahwa seleksi fitur *Information Gain* mampu mereduksi dimensi data sebesar 31,3% sekaligus meningkatkan performa klasifikasi jenis kelamin berbasis data kraniometrik.

ABSTRACT

Sex identification from human skulls is a crucial aspect of forensic anthropology; however, traditional methods still face limitations such as subjective assessment and inter-population variation. This study proposes the application of *Information Gain* as a *feature selection* technique and *Random Forest* as a classification algorithm for sex determination based on craniometric data. The dataset used is the Howells dataset consisting of 2,524 samples with 83 skull measurement features. *Feature selection* using *Information Gain* was performed with *threshold* values of 0.01, 0.05, and 0.09, followed by additional testing across a *threshold* range of 0.01 to 0.09. Model evaluation was conducted using 10-Fold *Cross Validation* with default *Random Forest* parameters. The results show that a *threshold* of 0.02 produced 57 selected features from the original 83, achieving the best performance with an *accuracy* of 87.40%, *precision* of 87.53%, *recall* of 87.40%, and *F1-score* of 87.41%. These results outperform the *baseline* model without *feature selection*, which achieved an *accuracy* of 86.57%. This study demonstrates that *Information Gain feature selection* can reduce data dimensionality by 31.3% while simultaneously improving sex classification performance based on craniometric data.

Penulis Koresponden

Email: iis.afrianty@uin-suska.ac.id

Cara sitasi IEEE::

N. A. Firana, I. Afrianty, N. Novriyanto, F. Yanto, "Application of *Information Gain* and *Random Forest* for Craniometric Sex Classification", *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 6, no. 2, p. 118-129, Juni 2026. doi:10.30811/ jaise.v6i2.9409

1. PENDAHULUAN

Antropologi forensik merupakan cabang ilmu yang mempelajari sisa-sisa tubuh manusia, seperti tulang, tengkorak, dan mumi, untuk tujuan identifikasi individu [1]. Bidang ini berperan penting dalam memperkirakan karakteristik biologis, seperti etnis, jenis kelamin, tinggi badan, dan usia dari individu yang belum diketahui identitasnya [2], [3]. Penentuan jenis kelamin menjadi salah satu tahapan penting dalam identifikasi forensik karena mampu mengurangi ruang pencarian identitas dan mendukung estimasi karakteristik biologis lainnya. [4], [5], [6]. Dalam praktiknya, identifikasi forensik banyak diterapkan pada kasus jenazah tanpa identitas, korban bencana alam, korban kebakaran, maupun jenazah yang ditemukan dalam kondisi tidak utuh [7]. Di antara berbagai bagian kerangka manusia, tulang tengkorak merupakan salah satu bagian yang paling sering digunakan karena memiliki karakteristik dimorfisme seksual yang jelas dan tingkat keawetan yang tinggi. Bahkan, tengkorak mampu memberikan tingkat akurasi hingga 90% dalam klasifikasi jenis kelamin setelah tulang panggul [8], [9], [10].

Metode tradisional dalam penentuan jenis kelamin umumnya menggunakan pendekatan morfologis dan metrik [10]. Meskipun relatif mudah dan cepat diterapkan, metode tersebut memiliki beberapa keterbatasan, seperti subjektivitas penilaian, ketergantungan pada kondisi kerangka, serta variasi karakteristik antar populasi yang dapat memengaruhi akurasi hasil [4], [11]. Selain itu, metode berbasis analisis DNA yang dikenal memiliki tingkat akurasi tinggi juga menghadapi kendala berupa biaya yang mahal, waktu analisis yang relatif lama, serta kemungkinan kegagalan ekstraksi DNA pada kondisi tulang yang terbakar atau mengalami kerusakan berat [12], [13]. Keterbatasan tersebut menunjukkan perlunya pendekatan alternatif yang lebih efisien, objektif, dan mampu menghasilkan klasifikasi yang akurat berdasarkan data pengukuran tulang tengkorak [14].

Perkembangan teknologi *machine learning* memberikan peluang baru dalam mengatasi permasalahan tersebut [15]. Dalam bidang antropologi forensik, *machine learning* mampu mempelajari pola dari data kraniometrik dan menghasilkan model klasifikasi dengan tingkat akurasi yang tinggi [16], [17]. Berbagai algoritma *machine learning*, seperti Artificial Neural Network (ANN), *Support Vector Machine* (SVM), dan *Random Forest* telah banyak diterapkan dalam klasifikasi data forensik. Ref. [18] menunjukkan bahwa *Random Forest* memiliki performa terbaik dibandingkan beberapa algoritma lainnya dengan akurasi mencapai 90,3%. Penelitian lain menunjukkan bahwa *Random Forest* mampu mencapai akurasi hingga 97,4% dalam estimasi jenis kelamin subdewasa [19]. Selain itu, Prabha et al. melaporkan bahwa *Random Forest* memiliki performa lebih baik dibandingkan *XGBoost* dalam penentuan jenis kelamin melalui analisis indeks mandibula dengan akurasi sebesar 97,20% [20]. Hasil tersebut menunjukkan bahwa *Random Forest* merupakan salah satu algoritma yang potensial untuk klasifikasi jenis kelamin pada data forensik.

Meskipun demikian, dataset kraniometrik umumnya memiliki jumlah fitur yang relatif besar sehingga dapat meningkatkan kompleksitas model dan waktu pembelajaran [14]. Oleh karena itu, diperlukan teknik seleksi fitur untuk mengurangi atribut yang tidak relevan sekaligus meningkatkan efisiensi dan akurasi klasifikasi. Salah satu metode seleksi fitur yang banyak digunakan adalah *Information Gain* , yang mampu mengukur tingkat relevansi setiap fitur terhadap kelas target [21]. Pada bidang antropologi forensik, penelitian oleh Tsawaabul Khair et al. menunjukkan bahwa kombinasi *Information Gain* dan *Backpropagation Neural Network* mampu meningkatkan akurasi klasifikasi jenis kelamin berbasis tengkorak dari 92,32% menjadi 93,91% [22]. Selain itu, penelitian pada domain lain juga menunjukkan bahwa *Information Gain* dapat meningkatkan performa klasifikasi, seperti pada klasifikasi ulasan pengguna aplikasi Discord [23] dan prediksi kanker payudara menggunakan *Random Forest* [24].

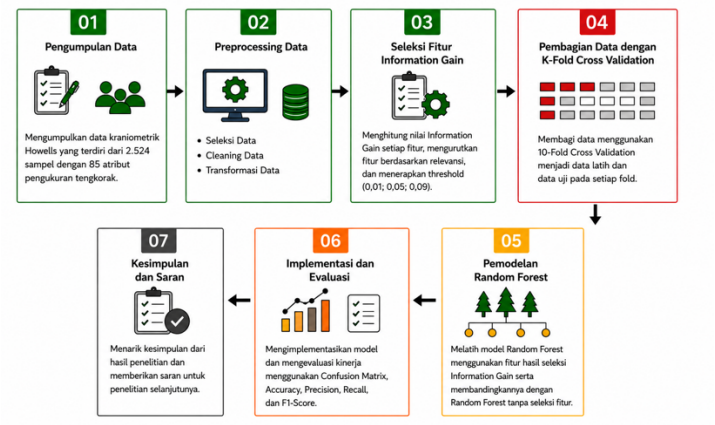
Keberhasilan berbagai penelitian telah membuktikan efektivitas *Random Forest* maupun *Information Gain* , penelitian yang mengombinasikan kedua metode tersebut pada dataset kraniometrik Howells untuk klasifikasi jenis kelamin masih terbatas. Penelitian sebelumnya pada dataset Howells lebih banyak berfokus pada penerapan *Information Gain* dengan *Backpropagation Neural Network*, sedangkan kajian mengenai pengaruh seleksi fitur *Information Gain* terhadap kinerja *Random Forest* pada dataset yang sama belum banyak dilaporkan. Selain memiliki jumlah fitur yang relatif besar, dataset Howells juga mencakup sampel dari berbagai populasi sehingga memiliki karakteristik data yang beragam. Oleh karena itu, belum diketahui secara jelas sejauh mana seleksi fitur *Information Gain* dapat meningkatkan efisiensi serta mempertahankan performa

klasifikasi jenis kelamin menggunakan algoritma *Random Forest*. Kondisi tersebut menunjukkan adanya research gap yang perlu diteliti lebih lanjut dalam penerapan metode *machine learning* pada data kraniometrik.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan penerapan *Information Gain* sebagai teknik seleksi fitur dan *Random Forest* sebagai algoritma klasifikasi untuk menentukan jenis kelamin berdasarkan data tulang tengkorak. Penelitian ini bertujuan untuk menganalisis pengaruh seleksi fitur *Information Gain* terhadap kinerja *Random Forest* serta mengevaluasi kemampuan model dalam mengklasifikasikan jenis kelamin menggunakan dataset kraniometrik Howells. Hasil penelitian diharapkan dapat meningkatkan efisiensi proses klasifikasi melalui reduksi fitur tanpa menurunkan performa model secara signifikan serta memberikan kontribusi dalam pengembangan metode berbasis *machine learning* pada bidang antropologi forensik.

2. METODE

Penelitian ini menerapkan metode *Information Gain* sebagai teknik seleksi fitur dan *Random Forest* sebagai metode klasifikasi. Tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian *Information Gain* dan *Random Forest* untuk Klasifikasi Jenis Kelamin Berdasarkan Data Kraniometrik

2.1 Pengumpulan Data

Penelitian ini menggunakan data sekunder yang berasal dari dataset Howells yang tersedia pada situs resmi Howells Craniometric Data Set. Dataset tersebut merupakan kumpulan data kraniometrik yang disusun oleh Dr. William W. Howells dan mencakup 2.524 individu dari berbagai populasi di dunia, yang terdiri atas 1.368 individu berjenis kelamin *male* dan 1.156 individu berjenis kelamin *female*. Data kraniometrik tersebut memuat 85 atribut linier hasil pengukuran tengkorak menggunakan kaliper, yang mencerminkan berbagai karakteristik morfometrik, seperti panjang, lebar, sudut, serta jari-jari tengkorak. Atribut-atribut tersebut digunakan untuk merepresentasikan variasi morfologi tengkorak yang berkaitan dengan jenis kelamin. Rincian mengenai fitur dan parameter pengukuran yang digunakan dalam penelitian ini disajikan pada Tabel 1 dan Tabel 2.

Tabel 1. Fitur Kraniometri Tulang Tengkorak

Kode	Fitur Kraniometri Tulang Tengkorak
GOL	Glabello-Occipital Length
NOL	Nasio-Occipital Length
BNL	Basion-Nasion Length
BBH	Basion-Bregma Height
XCB	Maximum Cranial Breadth
XFB	Maximum Frontal Breadth
ZYB	Bizygomatic Breadth
AUB	Biauricular Breadth
...	…
TBA	Thiobarbituric Acid

Tabel 2. Parameter Pengukuran Kraniometri Tengkorak

ID	Sex	PopNum	Population	GOL	NOL	BNL	BBH	XCB	XFB	ZYB	AUB	...	TBA
1	M	1	NORSE	189	185	100	135	143	120	133	119	…	0
2	M	1	NORSE	182	178	102	139	145	120	137	125	…	0
3	M	1	NORSE	191	187	102	123	140	114	134	125	…	0
...	…	…	…	…	…	…	…	…	…	…	…	…	…
3188	F	24	ANDAMAN	160	160	89	121	129	106	117	112	…	156
3189	F	24	ANDAMAN	172	170	92	118	137	110	114	106	…	167
3190	F	24	ANDAMAN	153	153	88	116	130	105	115	103	…	158

Berdasarkan Tabel 2, nilai ID tertinggi mencapai 3.190, namun jumlah sampel yang tersedia hanya 2.524 data. Hal ini menunjukkan adanya beberapa ID yang tidak memiliki data pengukuran, sehingga jumlah sampel aktual lebih kecil daripada rentang ID yang tercatat.

2.2 *Preprocessing Data*

Sebelum proses pemodelan dilakukan, data terlebih dahulu melalui tahap praproses untuk memastikan kualitas dan konsistensi data yang akan digunakan dalam klasifikasi [25]. Tahapan yang dilakukan meliputi:

a. *Seleksi Data:* Dataset awal terdiri dari 86 atribut dan satu label kelas. Atribut ID dihapus karena hanya berfungsi sebagai penomoran sampel tanpa nilai informatif. Atribut Population turut dihapus karena merupakan label demografis yang tidak merepresentasikan karakteristik morfometrik tengkorak secara langsung. Penyertaan atribut ini berpotensi menimbulkan *data leakage* serta menurunkan kemampuan generalisasi model, mengingat informasi asal populasi individu umumnya tidak tersedia dalam konteks identifikasi forensik. Setelah proses seleksi, diperoleh 83 atribut sebagai fitur dan satu atribut sebagai label kelas (Sex).
b. *Cleaning Data:* Tahap ini bertujuan untuk memeriksa keberadaan *missing value*, data duplikat, dan nilai nol pada dataset. Apabila ditemukan *missing value*, penanganan dilakukan melalui imputasi menggunakan nilai rata-rata (*mean*). Apabila terdapat data duplikat, baris yang berulang akan dihapus sehingga hanya menyisakan satu data unik. Adapun nilai nol yang teridentifikasi pada sejumlah fitur morfologi tidak diperlakukan sebagai kesalahan input, melainkan merepresentasikan bagian tulang yang tidak dapat diukur akibat kondisi fisik spesimen, sehingga nilai tersebut tetap dipertahankan dalam dataset.
c. *Transformasi Data:* Transformasi dilakukan pada atribut kelas (*Sex*) dengan mengubah label *Male* menjadi 1 dan *Female* menjadi 0 melalui proses *encoding*. Hasil transformasi menghasilkan dataset numerik yang siap digunakan pada tahap seleksi fitur *Information Gain* dan klasifikasi menggunakan *Random Forest*.

2.3 *Feature selection Information Gain*

Information Gain digunakan sebagai teknik seleksi fitur untuk menilai kontribusi setiap atribut dalam membedakan kelas target melalui pengurangan nilai ketidakpastian (*entropy*). Pada penelitian ini, nilai *Information Gain* dihitung untuk setiap atribut kraniometrik terhadap atribut kelas (Sex). Fitur dengan nilai *Information Gain* yang tinggi dipertahankan, sedangkan fitur dengan nilai rendah dieliminasi untuk mengurangi dimensi data dan meningkatkan performa klasifikasi. Proses seleksi fitur dilakukan melalui tahapan berikut:

a. Menghitung nilai *entropy* menggunakan Persamaan (1).

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \quad (1)$$

b. Menghitung nilai *Information Gain* setiap atribut menggunakan Persamaan (2).

$$Gain\ (S, A) = Entropy\ (S) - \sum_{values\ (A)} \frac{|Sv|}{|S|}\ Entropy\ (Sv) \quad (2)$$

Pada Persamaan (2), Gain(S,A) menunjukkan kemampuan atribut A dalam mengurangi ketidakpastian data, sedangkan Entropy(S) dan Entropy(Sv) masing-masing menyatakan tingkat ketidakpastian sebelum dan sesudah pemisahan data berdasarkan atribut tertentu.

2.4 Pembagian Data dengan *K-Fold Cross Validation*

K-Fold *Cross Validation* digunakan untuk membagi dataset menjadi data pelatihan dan data pengujian secara bergantian pada setiap iterasi. Metode ini menghasilkan evaluasi yang lebih stabil dan representatif

dibandingkan pembagian data secara langsung serta membantu mengurangi risiko *overfitting*. Pada penelitian ini digunakan 10-Fold *Cross Validation*, di mana sembilan fold digunakan sebagai data pelatihan dan satu fold sebagai data pengujian pada setiap iterasi [26].

2.5 *Pemodelan Random Forest*

Random Forest merupakan algoritma ensemble learning yang dibangun dari kumpulan *decision tree* menggunakan teknik *bootstrap* aggregation (*bagging*) dan random *feature selection* [27]. Pada penelitian ini, proses klasifikasi dilakukan melalui tahapan berikut:

a. Memasukkan data latih hasil pembagian menggunakan *10-Fold Cross Validation*.
b. Membentuk beberapa subset data pelatihan menggunakan teknik *bootstrap sampling* dengan pengambilan sampel secara acak disertai pengembalian (*sampling with replacement*).
c. Membangun sejumlah *decision tree* dengan pemilihan fitur secara acak dan penentuan atribut terbaik menggunakan *Gini Index* pada Persamaan (3).

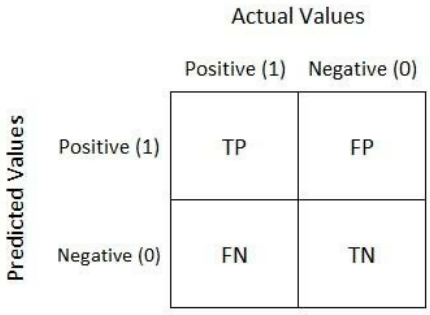
$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2 \quad (3)$$

d. Melakukan prediksi terhadap data uji menggunakan seluruh *decision tree* yang terbentuk.
e. Menggabungkan hasil prediksi menggunakan metode *majority voting* untuk memperoleh kelas akhir

2.6 Implementasi dan Evaluasi

Implementasi penelitian dilakukan menggunakan bahasa pemrograman Python pada platform Google Colab. Proses komputasi didukung oleh perangkat keras berupa Apple M4 Chip yang terdiri atas CPU 10-Core dan GPU 10-Core, RAM sebesar 16 GB, serta media penyimpanan SSD berkapasitas 256 GB.

Evaluasi dilakukan dengan membandingkan kinerja algoritma *Random Forest* tanpa seleksi fitur dan *Random Forest* yang dikombinasikan dengan *Information Gain* pada nilai *threshold* 0,01; 0,05; dan 0,09. Seluruh eksperimen menggunakan parameter standar (*default*) pada algoritma *Random Forest* serta menerapkan metode 10-Fold *Cross Validation* untuk menghasilkan evaluasi yang lebih stabil dan representatif. Kinerja model diukur menggunakan *Confusion Matrix* yang terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan nilai-nilai tersebut, dihitung metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk menilai kemampuan model dalam mengklasifikasikan jenis kelamin. Ilustrasi *Confusion Matrix* yang digunakan pada penelitian ini ditunjukkan pada Gambar 2.



Gambar 2. *Confusion Matrix*

Dari *confusion matrix*, sejumlah metrik evaluasi dapat dihitung seperti berikut :

a. *Accuracy*: mengukur proporsi prediksi yang diklasifikasikan dengan benar terhadap seluruh data uji. Perhitungan accuracy menggunakan Persamaan (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

b. *Precision*: mengukur tingkat ketepatan prediksi pada kelas positif yang dihasilkan oleh model. Perhitungan precision menggunakan Persamaan (5).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

c. *Recall*: mengukur kemampuan model dalam mengenali seluruh data yang termasuk ke dalam kelas positif. Perhitungan recall menggunakan Persamaan (6).

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

d. *F1-score*: merupakan rata-rata harmonis antara precision dan recall yang digunakan untuk memberikan penilaian yang lebih seimbang terhadap performa model. Perhitungan F1-Score menggunakan Persamaan (7).

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{\text{Presisi} + Recall} \quad (7)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berjumlah 2.524 sampel tulang tengkorak, terdiri atas 1.368 sampel *male* dan 1.156 sampel *female*, dengan 85 fitur pengukuran sebagai dasar proses klasifikasi jenis kelamin.

3.2 *Preprocessing Data*

a. *Seleksi Data:* Pada langkah ini, dilakukan penghapusan fitur ID dan Population, sehingga fitur yang tersisa berjumlah 83 fitur. Hasil dari seleksi data tersaji dalam Gambar 3.

```
Shape dataset awal: (2524, 86)

Kolom dataset:
['ID', 'Sex', 'PopNum', 'Population', 'GOL', 'NOL', 'BNL', 'BBH', 'XCB', 'XFB', 'ZYB', 'AUB', 'WCB', 'ASB', 'BPL', 'NPH', 'NLH',
```

	ID	Sex	PopNum	Population	GOL	NOL	BNL	BBH	XCB	XFB	...	FRA	PAA	OCA	RFA	RPA	ROA	BSA	SBA	SLA	TBA
0	1	M	1	NORSE	189	185	100	135	143	120	...	134	133	117	0	0	0	0	0	0	0
1	2	M	1	NORSE	182	178	102	139	145	120	...	128	134	119	0	0	0	0	0	0	0
2	3	M	1	NORSE	191	187	102	123	140	114	...	129	137	111	0	0	0	0	0	0	0
3	4	M	1	NORSE	191	188	100	127	141	123	...	128	135	108	0	0	0	0	0	0	0
4	5	M	1	NORSE	178	177	97	128	138	117	...	133	130	111	0	0	0	0	0	0	0

5 rows × 86 columns

Gambar 3. Hasil Seleksi Data

b. *Cleaning Data:* Tahap ini bertujuan untuk memeriksa keberadaan data duplikat, *missing value*, dan nilai nol pada dataset. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung data duplikat maupun *missing value*. Namun, teridentifikasi sebanyak 7.320 nilai nol yang tersebar pada 11 fitur morfologi, yaitu BRR, LAR, OSR, BAR, BSA, RFA, RPA, ROA, SLA, SBA, dan TBA, serta 38 nilai nol pada fitur GLS. Kemunculan nilai nol pada fitur-fitur tersebut tidak diperlakukan sebagai kesalahan input data, melainkan mencerminkan kondisi di mana bagian tulang tertentu tidak dapat diukur akibat keterbatasan fisik spesimen. Mengacu pada pertimbangan tersebut, seluruh nilai nol tetap dipertahankan dalam dataset tanpa dilakukan imputasi maupun penghapusan. Hasil proses *cleaning* ditunjukkan pada Gambar 4.

```
# ============================================================
# CELL 5 – CLEANING (cek missing value & duplikat)
# ============================================================
print("Jumlah missing value per kolom:")
mv = df.isnull().sum()
print(mv[mv > 0] if mv.sum() > 0 else "Tidak ada missing value")

print(f"\nJumlah duplikat : {df.duplicated().sum()}")
print(f"Jumlah nilai 0  : {(df[df.columns.difference(['Sex'])] == 0).sum().sum()}")

# Hapus baris dengan missing value jika ada
df = df.dropna()
print(f"\nShape setelah cleaning: {df.shape}")
```

```
Jumlah missing value per kolom:
Tidak ada missing value

Jumlah duplikat : 0
Jumlah nilai 0  : 7320

Shape setelah cleaning: (2524, 84)
```

(a)

```
import pandas as pd

df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/DATASET/data_tulang_tengkorak.csv')

# Cek kolom mana saja yang punya nilai 0
zero_counts = (df == 0).sum()
zero_cols = zero_counts[zero_counts > 0].sort_values(ascending=False)

print("Kolom yang memiliki nilai 0:")
print(zero_cols.to_string())
print(f"\nTotal nilai 0 keseluruhan: {zero_counts.sum()}")
```

```
Kolom yang memiliki nilai 0:
BRR    662
LAR    662
OSR    662
BAR    662
BSA    662
RFA    662
RPA    662
ROA    662
SLA    662
SBA    662
TBA    662
GLS     38

Total nilai 0 keseluruhan: 7320
```

(b)

Gambar 4. Cleaning Data Untuk (a) Pengecekan Jumlah dan (b) Menampilkan Kolom Nilai 0

c. *Transfomasi Data:* Tahap ini dilakukan dengan mengonversi atribut *Sex* ke dalam bentuk numerik melalui proses *encoding*, yaitu *Male* menjadi 1 dan *Female* menjadi 0. Hasil transformasi data disajikan pada Gambar 5.

```
Distribusi label setelah transformasi:
Sex
1    1368
0    1156
Name: count, dtype: int64

Total data : 2524
Pria   (1) : 1368
Wanita (0) : 1156

Jumlah fitur awal : 83
Jumlah sampel     : 2524
```

```
df.head()
```

	Sex	PopNum	GOL	NOL	BNL	BBH	XCB	XFB	ZYB	AUB	...	FRA	PAA	OCA	RFA	RPA	ROA	BSA	SBA	SLA	TBA
0	1	1	189	185	100	135	143	120	133	119	...	134	133	117	0	0	0	0	0	0	0
1	1	1	182	178	102	139	145	120	137	125	...	128	134	119	0	0	0	0	0	0	0
2	1	1	191	187	102	123	140	114	134	125	...	129	137	111	0	0	0	0	0	0	0
3	1	1	191	188	100	127	141	123	135	127	...	128	135	108	0	0	0	0	0	0	0
4	1	1	178	177	97	128	138	117	129	121	...	133	130	111	0	0	0	0	0	0	0

5 rows × 84 columns

Gambar 5. Hasil Transformasi Data Pada Fitur Sex

3.3 *Feature selection Information Gain*

Proses seleksi fitur menggunakan *Information Gain* diawali dengan perhitungan nilai *entropy* pada atribut *Sex* menggunakan Persamaan (2). Hasil perhitungan *entropy* ditampilkan pada Gambar 6.

```
Entropy Dataset (Sex) : 0.9949
```

Gambar 6. Hasil Perhitungan Entropy

Setelah nilai *entropy* diperoleh, langkah berikutnya adalah menghitung nilai *Information Gain* setiap fitur menggunakan Persamaan (3). Proses ini menghasilkan nilai relevansi masing-masing fitur terhadap variabel target. Selanjutnya, nilai *Information Gain* diurutkan dari yang tertinggi hingga terendah berdasarkan tingkat kontribusinya dalam proses klasifikasi. Hasil perhitungan *Information Gain* disajikan pada Tabel 3.

Tabel 3. Hasil Perhitungan *Information Gain*

No.	Fitur	*Information Gain*	No.	Fitur	*Information Gain*
1.	ZYB	0.269319	43.	LAR	0.035887
2.	JUB	0.222355	44.	PAF	0.034393
3.	MDH	0.201457	45.	SIA	0.032451
4.	GOL	0.159275	46.	OSR	0.031984
5.	MDB	0.156351	47.	NLB	0.029678
6.	FMB	0.153848	48.	FRA	0.028979
7.	ZMB	0.146428	49.	DKB	0.028047
8.	AUB	0.137967	50.	SSS	0.024919
9.	XML	0.136636	51.	SIS	0.024836
10.	MAB	0.132449	52.	NDS	0.024680
11.	SOS	0.132075	53.	OCC	0.023322
12.	NOL	0.130684	54.	NAS	0.022706
13.	NAR	0.122966	55.	STB	0.022612
14.	EKB	0.121212	56.	PopNum	0.022083
15.	BNL	0.115728	57.	BAR	0.021753
16.	GLS	0.112892	58.	OCF	0.018513
17.	BBH	0.112603	59.	PAS	0.017887
18.	FRC	0.106476	60.	OBH	0.010970
19.	NPH	0.104835	61.	NAA	0.009345
20.	NLH	0.102624	62.	OCS	0.009108
21.	VRR	0.101358	63.	BAA	0.008471
22.	DKR	0.095538	64.	SLA	0.008079
23.	AVR	0.094018	65.	NDA	0.007833
24.	SSR	0.090389	66.	BSA	0.007352
25.	WCB	0.084684	67.	BRA	0.005312
26.	ASB	0.084528	68.	RPA	0.004925
27.	PRR	0.082207	69.	BBA	0.004553
28.	BRR	0.081337	70.	TBA	0.004413
29.	FRF	0.079633	71.	RFA	0.004308
30.	FMR	0.076601	72.	SBA	0.004308
31.	EKR	0.075896	73.	ROA	0.004308
32.	XCB	0.073837	74.	FRS	0.003578

33.	ZOR	0.072184	75.	DKA	0.003491
34.	OBB	0.070643	76.	DKS	0.002898
35.	PAC	0.070251	77.	NFA	0.002867
36.	XFB	0.067826	78.	PRA	0.002811
37.	ZMR	0.066576	79.	WNB	0.002807
38.	FOL	0.064408	80.	OCA	0.001794
39.	IML	0.060721	81.	NBA	0.001705
40.	WMH	0.056457	82.	PAA	0.000736
41.	BPL	0.054390	83.	SSA	0.000519
42.	MLS	0.042700			

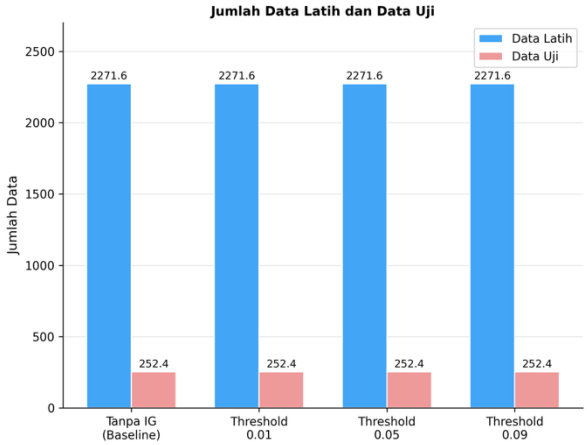
Berdasarkan nilai *Information Gain* yang diperoleh pada Tabel 3, dilakukan proses seleksi fitur menggunakan tiga nilai *threshold*, yaitu 0,01; 0,05; dan 0,09. Hasil seleksi menunjukkan bahwa *threshold* 0,01 menghasilkan 60 fitur terpilih, *threshold* 0,05 menghasilkan 41 fitur, sedangkan *threshold* 0,09 menghasilkan 24 fitur. Semakin besar nilai *threshold* yang digunakan, semakin sedikit fitur yang dipertahankan karena hanya fitur dengan nilai *Information Gain* yang melebihi batas tersebut yang akan dipilih. Ringkasan hasil seleksi fitur untuk setiap *threshold* ditampilkan pada Tabel 4.

Tabel 4. Jumlah Fitur Terpilih Berdasarkan Nilai *Threshold Information Gain*

Threshold	Jumlah Fitur Terpilih
0,01	60
0,05	41
0,09	24

3.4 Pembagian Data dengan *K-Fold Cross Validation*

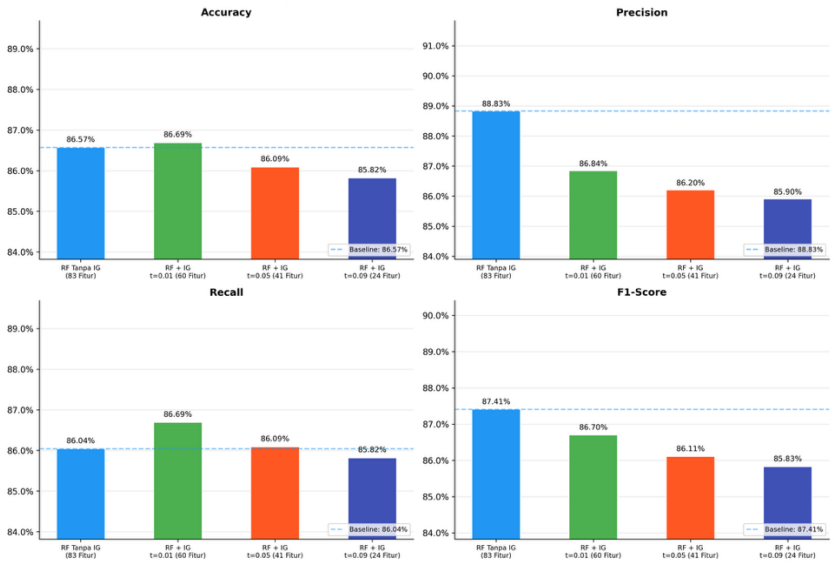
Untuk memperoleh evaluasi model yang lebih representatif, penelitian ini menerapkan metode 10-*Fold Cross Validation*. Dataset dibagi menjadi sepuluh subset yang relatif seimbang, di mana sembilan subset digunakan sebagai data pelatihan dan satu subset digunakan sebagai data pengujian pada setiap iterasi. Proses ini dilakukan secara bergantian hingga seluruh subset memperoleh kesempatan menjadi data uji satu kali. Rincian pembagian data pada setiap *fold* disajikan pada Gambar 5.



Gambar 7. Rincian Pembagian Data Pada K=10

3.5 Implementasi dan Evaluasi

Pengujian dilakukan melalui dua skenario, yaitu Random Forest tanpa seleksi fitur sebagai model pembanding (*baseline*) dan *Random Forest* yang dikombinasikan dengan seleksi fitur *Information Gain*. Kedua skenario dievaluasi menggunakan metode 10-*Fold Cross Validation* dengan parameter standar (*default*) dari pustaka *Scikit-Learn*, yaitu *n_estimators*=100, *max_depth*=None, *max_features*='sqrt', *min_samples_split*=2, *min_samples_leaf*=1, dan *bootstrap=True*. Evaluasi performa dilakukan menggunakan metrik Accuracy, Precision, Recall, dan F1-Score. Hasil pengujian masing-masing skenario ditampilkan pada Tabel 6.

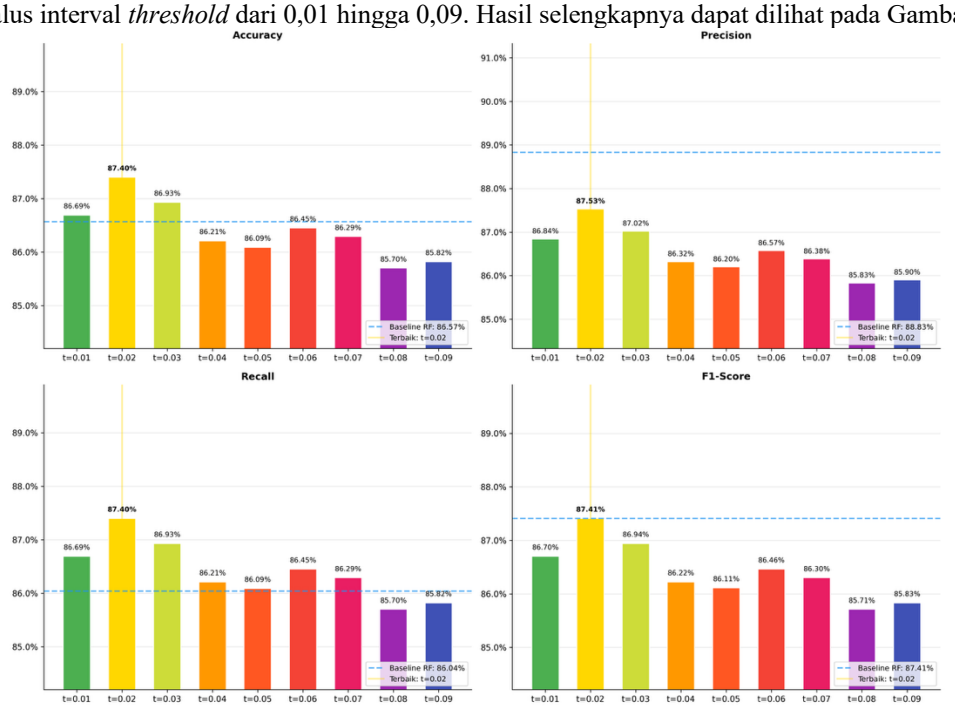


Gambar 8. Hasil Pengujian

Berdasarkan Gambar 8, kedua skenario menunjukkan hasil yang berbeda. Skenario pertama merupakan model *baseline* yaitu *Random Forest* tanpa seleksi fitur yang menggunakan seluruh 83 fitur tersedia. Hasil pengujian menunjukkan nilai *accuracy* sebesar 86,57%, *precision* 88,83%, *recall* 86,04%, dan *F1-score* 87,41%. Model ini menjadi acuan perbandingan untuk mengukur sejauh mana seleksi fitur *Information Gain* memberikan pengaruh terhadap performa klasifikasi.

Skenario kedua menggunakan kombinasi *Random Forest* dengan seleksi fitur *Information Gain* pada tiga variasi nilai *threshold* yaitu 0,01, 0,05, dan 0,09. Hasil pengujian menunjukkan bahwa *threshold* 0,01 dengan 60 fitur terpilih menghasilkan performa terbaik dibandingkan dua *threshold* lainnya, dengan nilai *accuracy* 86,69%, *precision* 86,84%, *recall* 86,69%, dan *F1-score* 86,70%. Peningkatan performa ini menunjukkan bahwa seleksi fitur menggunakan *Information Gain* mampu menyaring fitur-fitur yang kurang relevan sehingga model dapat bekerja lebih optimal meskipun dengan jumlah fitur yang lebih sedikit. Sebaliknya, *threshold* yang terlalu tinggi seperti 0,09 justru menurunkan performa karena terlalu banyak fitur yang dieliminasi sehingga informasi penting ikut terbuang.

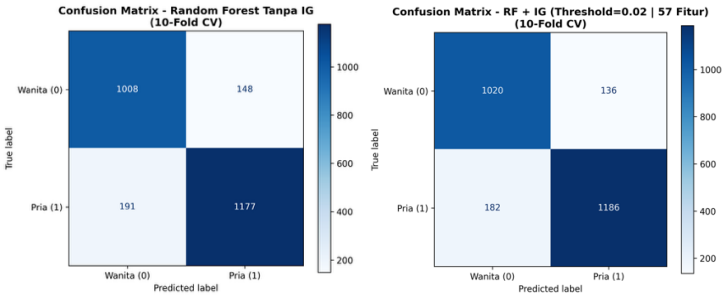
Untuk mendapatkan nilai *threshold* yang lebih optimal, dilakukan pengujian tambahan dengan memperhalus interval *threshold* dari 0,01 hingga 0,09. Hasil selengkapnya dapat dilihat pada Gambar 9.



Gambar 9. Hasil Pengujian Tambahan Variasi *Threshold*

Berdasarkan hasil pengujian pada Gambar 9, *threshold* 0,02 dengan 57 fitur terpilih menghasilkan performa terbaik dibandingkan seluruh variasi yang diuji. Model tersebut memperoleh nilai Accuracy sebesar 87,40%, Precision sebesar 87,53%, Recall sebesar 87,40%, dan F1-Score sebesar 87,41%. Secara keseluruhan, performa terbaik diperoleh pada skenario Random Forest dengan seleksi fitur Information Gain menggunakan *threshold* 0,02. Dari 83 fitur awal, hanya 57 fitur yang dipertahankan sehingga terjadi reduksi dimensi sebesar 31,3%. Selain itu, model juga menunjukkan peningkatan pada seluruh metrik evaluasi dibandingkan model *baseline*.

Untuk memberikan gambaran yang lebih rinci mengenai hasil klasifikasi, dilakukan analisis terhadap *Confusion Matrix*dari model *baseline* dan model terbaik yang ditampilkan pada Gambar 10.



Gambar 10. Perbandingan Confusion Matrix

Berdasarkan Gambar 10, penerapan Information Gain memberikan perbaikan terhadap performa klasifikasi. Pada model Random Forest tanpa seleksi fitur, model berhasil mengklasifikasikan 1.008 data perempuan secara benar (*True Negative*) dan 1.177 data laki-laki secara benar (*True Positive*). Namun demikian, masih terdapat 148 data perempuan yang salah diklasifikasikan sebagai laki-laki (*False Positive*) serta 191 data laki-laki yang salah diklasifikasikan sebagai perempuan (*False Negative*).

Setelah seleksi fitur Information Gain diterapkan dengan *threshold* 0,02, jumlah *True Negative* meningkat menjadi 1.020 dan *True Positive* meningkat menjadi 1.186. Di sisi lain, jumlah *False Positive* menurun menjadi 136 dan *False Negative*berkurang menjadi 182. Hasil ini menunjukkan bahwa eliminasi fitur yang kurang relevan mampu membantu model dalam mempelajari pola data secara lebih efektif sehingga meningkatkan kemampuan klasifikasi jenis kelamin berdasarkan data kraniometrik.

4. KESIMPULAN

Penelitian ini telah berhasil menerapkan *Information Gain* sebagai teknik seleksi fitur dan *Random Forest* sebagai algoritma klasifikasi untuk menentukan jenis kelamin berdasarkan data kraniometrik pada dataset Howells yang terdiri atas 2.524 sampel dan 83 fitur. Pengujian dilakukan melalui dua skenario, yaitu *Random Forest* tanpa seleksi fitur sebagai model pembanding (*baseline*) dan *Random Forest* yang dikombinasikan dengan seleksi fitur *Information Gain* menggunakan *threshold* 0,01; 0,05; dan 0,09. Pengujian lanjutan menunjukkan bahwa *threshold* 0,02 menghasilkan kinerja paling optimal dengan 57 fitur terpilih, memperoleh nilai *Accuracy* sebesar 87,40%, *Precision* sebesar 87,53%, *Recall* sebesar 87,40%, dan *F1-Score* sebesar 87,41%. Peningkatan sebesar 0,83% terhadap model *baseline* bersifat marginal, yang dapat dijelaskan karena *Random Forest* secara inheren sudah memiliki mekanisme seleksi fitur internal melalui *random feature selection*. Fitur-fitur terpilih pada *threshold* 0,02, seperti ZYB, JUB, dan MDH, merepresentasikan dimensi wajah yang memiliki dimorfisme seksual paling nyata secara biologis sehingga memperkuat kemampuan klasifikasi model. Sebaliknya, *threshold* yang terlalu tinggi menyebabkan eliminasi berlebihan terhadap fitur-fitur yang secara kolektif berkontribusi pada diversitas antar *decision tree*, sehingga performa model menurun. Kontribusi utama penelitian ini lebih ditekankan pada efisiensi komputasi melalui reduksi dimensi sebesar 31,3% tanpa mengorbankan akurasi secara berarti. Untuk penelitian selanjutnya, disarankan untuk mempertimbangkan metode seleksi fitur lain seperti *Recursive Feature Elimination* (RFE) dan *Mutual Information*, serta optimasi hiperparameter *Random Forest*. Pengujian pada dataset populasi yang lebih beragam, khususnya Asia Tenggara, juga perlu dilakukan untuk meningkatkan generalisasi model dalam konteks antropologi forensik.

REFERENSI

[1] G. Mason and H. Yusuf, “Physics and Geology as a crime solving science,” *Criminology*, pp. 2162–2175, 2024, [Online]. Available: https://jicnusantara.com/index.php/jicn

[2] F. R. Amalia *et al.*, “Identification of Forensic Odontology in Investigation Case of Human Skeletal Findings: Case Report,” vol. 06, no. 04, pp. 996–1006, 2025, doi:

10.37899/journallamedihealtico.v6i4.
[3] M. Zhang, "The Application of Forensic Imaging to Sex Estimation: Focus on Skull and Pelvic Structures," *Perspect. Leg. Forensic Sci.*, vol. 1, no. 1, pp. 10005–10005, 2024, doi: 10.35534/plfs.2024.10005.
[4] I. Afrianty, D. Nasien, and H. Haron, "Performance Analysis of *Support Vector Machine* in Sex Classification of The Sacrum Bone in Forensic Anthropology," *J. Tek. Inform.*, vol. 15, no. 1, pp. 63–72, Jun. 2022, doi: 10.15408/jti.v15i1.25254.
[5] N. S. Appel and H. J. H. Edgar, "A Pilot Study of Age Estimation and Cause of Death: Insights into Skeletal Aging," *Forensic Sci.*, vol. 4, no. 4, pp. 508–522, Oct. 2024, doi: 10.3390/forensicsci4040034.
[6] L. Sugara, F. Yanti, and Hanif, "Jurnal Kesehatan Dan Ilmu Kedokteran (JUKIK)," *J. Kesehat. dan ilmu Kedokt.*, vol. 06, no. 3, pp. 8–24, 2024.
[7] T. Suryadi, M. J. Ramadhanif, R. P. Sari, A. Wulandari, and F. Kamila, "Identifikasi Pada Jenazah Yang Ditemukan Di Pinggir Pantai," *Indones. J. Leg. Forensic Sci.*, vol. 11, no. 2, p. 112, 2021, doi: 10.24843/ijlfs.2021.v11.i02.p06.
[8] Arthy, R. Goel, and M. Sreenivas, "Determination of sex by osteometry of third metatarsal," *Indian J. Forensic Med. Toxicol.*, vol. 14, no. 3, pp. 1–6, 2020, doi: 10.37506/ijfmt.v14i3.10315.
[9] F. Curate, "The Estimation of Sex of Human Skeletal Remains in the Portuguese Identified Collections: History and Prospects," *Forensic Sci.*, vol. 2, no. 1, pp. 272–286, Mar. 2022, doi: 10.3390/forensicsci2010021.
[10] M. J. S. Mota *et al.*, "Enhancing sex determination in forensic anthropology: A comparative analysis of cranial measurements using artificial neural network," *Forensic Sci. Int. Reports*, 2025, doi: 10.1016/j.fsir.2025.100422.
[11] X. Wang, G. Liu, Q. Wu, Y. Zheng, F. Song, and Y. Li, "Sex estimation techniques based on skulls in forensic anthropology: A scoping review," *PLoS One*, vol. 19, no. 12, pp. 1–22, 2024, doi: 10.1371/journal.pone.0311762.
[12] D. Nasien, M. H. Adiya, I. Afrianty, N. A. Ali, A. A. Samah, and Y. Rahayu, "Determination of Sex and Race in Forensic Anthropology: A Comparison of Artificial Neural Network and *Support Vector Machine*," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, Sep. 2021, pp. 51–55. doi: 10.1109/IC2IE53219.2021.9649182.
[13] P. Mesejo, R. Martos, Ó. Ibáñez, and J. Novo, "applied sciences A Survey on Artificial Intelligence Techniques for Biomedical Image Analysis in Skeleton-Based Forensic Human Identification," 2020, doi: 10.3390/app10144703.
[14] S. S. Rahayu, I. Afrianty, E. Budianita, and F. Syafria, "Klasifikasi Tulang Tengkorak Berdasarkan Jenis Kelamin dalam Antropologi Forensik Menggunakan Metode *Support Vector Machine*," *INOVTEK Polbeng - Seri Inform.*, vol. 9, no. 1, pp. 243–256, 2024, doi: 10.35314/isi.v9i1.4046.
[15] E. Nikita and P. Nikitas, "On the use of *machine learning* algorithms in forensic anthropology," *Leg. Med.*, vol. 47, p. 101771, Nov. 2020, doi: 10.1016/j.legalmed.2020.101771.
[16] E. Faisal and T. L. Rogers, "A review of the literature on the applications of *machine learning* in forensic anthropology," *Forensic Sci. Int.*, vol. 376, p. 112579, Nov. 2025, doi: 10.1016/j.forsciint.2025.112579.
[17] W. Yang, M. Zhou, P. Zhang, G. Geng, X. Liu, and H. Zhang, "Skull Sex Estimation Based on Wavelet Transform and Fourier Transform," *Biomed Res. Int.*, vol. 2020, no. 1, Jan. 2020, doi: 10.1155/2020/8608209.
[18] S. Knecht, F. Santos, Y. Ardagna, V. Alunni, P. Adalian, and L. Nogueira, "Sex estimation from long bones: a *machine learning* approach," *Int. J. Legal Med.*, vol. 137, no. 6, pp. 1887–1895, Nov. 2023, doi: 10.1007/s00414-023-03072-4.
[19] M. Stephanie Cole, "Developing Subadult Sex Estimation Standards Using Adult Morphological Sex Traits and an Ontogenetic Approach," 2022.
[20] P. S. Prabha, A. Ganesan, K. C. Lakshmi, and A. J. Murugan, "Sex determination through analysis of mandibular indices using lateral cephalogram: An Artificial intelligence diagnostics," *Discov. Artif. Intell.*, vol. 5, no. 1, 2025, doi: 10.1007/s44163-025-00371-0.
[21] I. K. Hasan, R. Resmawan, and J. Ibrahim, "Perbandingan K-Nearest Neighbor dan *Random Forest* dengan Seleksi Fitur *Information Gain* untuk Klasifikasi Lama Studi Mahasiswa," *Indones. J. Appl. Stat.*, vol. 5, no. 1, p. 58, May 2022, doi: 10.13057/ijas.v5i1.58056.
[22] N. Tsawaabul Khair, I. Afrianty, F. Syafria, E. Budianita, and S. Kurnia Gusti, "Penerapan *Information Gain* Untuk Seleksi Fitur Pada Klasifikasi Jenis Kelamin Tulang Tengkorak Menggunakan Backpropagation," *Media Online)*, vol. 5, no. 4, pp. 666–678, 2025, doi: 10.47065/bulletincsr.v5i4.637.
[23] S. N. Salsabila, B. N. Sari, and R. Mayasari, "Klasifikasi Ulasan Pengguna Aplikasi Discord

Menggunakan Metode *Information Gain* Dan Naïve Bayes Classifier," *INFOTECH J.*, vol. 9, no. 2, pp. 383–392, Jul. 2023, doi: 10.31949/infotech.v9i2.6277.

[24] A. Demos, "Penerapan Metode *Random Forest* Dan *Information Gain* Untuk Prediksi Kanker Payudara," 2023, [Online]. Available: https://repository.ithb.ac.id/id/eprint/275/

[25] M. A. N. Anargya, W. Ghozi, and F. A. Rafrastara, "Optimizing IoV Attack Detection Using Random Under Sampling Techniques," *J. Inform. J. Pengemb. IT*, vol. 10, pp. 11–19, 2025, doi: 10.30591/jpit.v10i1.8034.

[26] V. W. Lumumba, D. Kiprotich, M. Lemasulani Mpaine, N. Grace Makena, and M. Daniel Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in *Machine learning* Models," *SSRN Electron. J.*, vol. 13, no. 5, pp. 127–137, 2025, doi: 10.2139/ssrn.5266507.

[27] A. Sunyoto and H. Al Fatta, "Klasifikasi Penyakit Jantung Menggunakan *Random Forest* Classifier," *J. Sist. Komput. dan Kecerdasan Buatan*, vol. VII, no. September, pp. 31–40, 2023.

[28] M. R. Adriana, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi *Random Forest* Dan Svm Pada Analisis Sentimen Psbb," *J. Inform. UPGRIS*, vol. 7, no. 1, pp. 36–40, 2021, [Online]. Available: https://doi.org/10.26877/jiu.v7i1.7099