Klasifikasi Material Lingkungan dengan Machine Learning: Membandingkan algoritma Random Forest, Decision Tree, dan SVM

Ellen Proborini*[1], Fajar Husain Asy'ari[2], Ratna Herawati[3], Bijanto[4]
[1,2,4]Sekolah Tinggi Teknik Pati
[3]Universitas Dharma AUB Surakarta
e-mail: [1]ellenproborini@gmail.com, [1]fajarhusain@sttp.ac.id, [3]ratna.herawati@stmik-aub.ac.id,
[4]biyantokakoi@gmail.com
*Penulis Korespondensi

Diterima: 21 Juni 2026; Direvisi: 18 Oktober 2026; Disetujui: 22 Oktober 2026

Abstrak

Tanah, air, batuan, mineral maupun logam memiliki karakter material yang hampir serupa, sehingga perlu diklasifikasi. Proses klasifikasi dilakukan dengan mengembangkan dan menguji model menggunakan Decision Tree, Random Forest, dan Support Vector Machine (SVM). Jenis material yang diklasifikasi yaitu Air, Metal, Mineral, Organic, Rock, Sediment, Soil, dan Water, dataset diambil dari Earth Intelligence Dataset (EID) dengan 5.000 sampel dan 42 fitur. Dataset berisi data komposisi kimia, sifat fisik dan data lingkungan, pem-prosesan data dengan imputasi missing value, normalisasi Z-score dan encoding label. Data dipisah 80:20, proses evaluasi menggunakan akurasi, recall, F1-score, sedangkan pengujian dengan cross validation. Hasil penelitian mendapatkan hasil akurasi 98% untuk Decision Tree, 97% Random Forest dan SVM 94,60%. Kelas mineral dan batu memiliki kesamaan karakteristik geologis sehingga proses klasifikasi lebih rumit. Terdapat 5 fitur penting dalam proses klasifikasi ini yaitu, densitas, kadar oksigen, kadar magnesium, sifat termal, dan kadar kalsium. Diketahui bahwa Decision Tree mendapat nilai akurasi tertinggi, dengan standard deviation 0,0036, yang menandakan stabilitas model yang baik. Penelitian berkontribusi pada literatur dengan kombinasi beberapa algoritma yang belum banyak dieksplorasi dalam proses identifikasi material lingkungan berbasis komposisi kimia dan sifat fisik.

Kata kunci: *klasifikasi material lingkungan, Random Forest, Decision Tree, SVM, machine learning*

Abstract

Soil, air, rocks, minerals, and metals have almost similar material characteristics, so they need to be classified. The classification process is carried out by developing and testing models using Decision Tree, Random Forest, and Support Vector Machine (SVM). The types of materials classified are Air, Metal, Mineral, Organic, Rock, Sediment, Soil, and Water. The dataset is taken from the Earth Intelligence Dataset (EID) with 5,000 samples and 42 features. The dataset contains chemical composition data, physical properties, and environmental data. Data processing uses missing value imputation, Z-score normalization, and coding labels. The data is split 80:20, the evaluation process uses accuracy, recall, and F1-score, while testing uses cross-validation. The results of the study obtained 98% accuracy for Decision Tree, 97% for Random Forest, and 94.60% for SVM. Mineral and rock classes have similar geological characteristics, making the classification process more complicated. There are 5 important features in this classification process, namely density, oxygen content, magnesium content, thermal properties, and calcium content. It is known that the Decision Tree achieved the highest accuracy value, with a standard deviation of 0.0036, indicating good model stability. This research contributes to the literature by providing several combinations of algorithms that have not been widely explored in the process of identifying environmental materials based on chemical composition and physical properties.

Keywords: *environmental material classification, random forest, decision tree, svm, machine learning*

I. PENDAHULUAN

This is an open access article under the CC–BY-SA license.

Perkembangan *Artificial Intelligent* dalam beberapa waktu terakhir membuka peluang besar dalam berbagai sektor, termasuk ilmu lingkungan dan material *sains* [1]. Kemampuan untuk identifikasi dan klasifikasi material lingkungan seperti tanah, air, logam, mineral, batuan, sedimen, dan material organik secara akurat dan otomatis menjadi kebutuhan mendasar dalam mendukung beberapa observasi sumber daya, pengelolaan limbah, serta mitigasi pencemaran [2], [3]. Setiap material memiliki komposisi unsur kimia, sifat fisika-kimia, serta respons terhadap faktor lingkungan yang khas. Karakteristik unik material dapat didapatkan dan dipelajari oleh algoritma *machine learning* dalam proses membangun sistem klasifikasi yang andal [4]. *Dataset multidomain* seperti *Earth Intelligence Dataset* (*EID*) yang dipublikasikan oleh Sumon [5] menyediakan data terstruktur, dimana data menghubungkan komposisi material (unsur H, O, N, C, logam), sifat fisik (pH, *densitas*, *konduktivitas*), dan konteks lingkungan (suhu, kelembaban, lokasi). *Dataset* ini dirancang khusus untuk mendukung berbagai tugas AI, mulai dari klasifikasi material, prediksi kualitas air, pemodelan kesuburan tanah, hingga estimasi kemurnian logam [5]. Keberadaan *dataset* terintegrasi seperti *EID* menjadi kunci dalam mengembangkan dan menguji performa berbagai algoritma *machine learning* pada kasus nyata lingkungan.

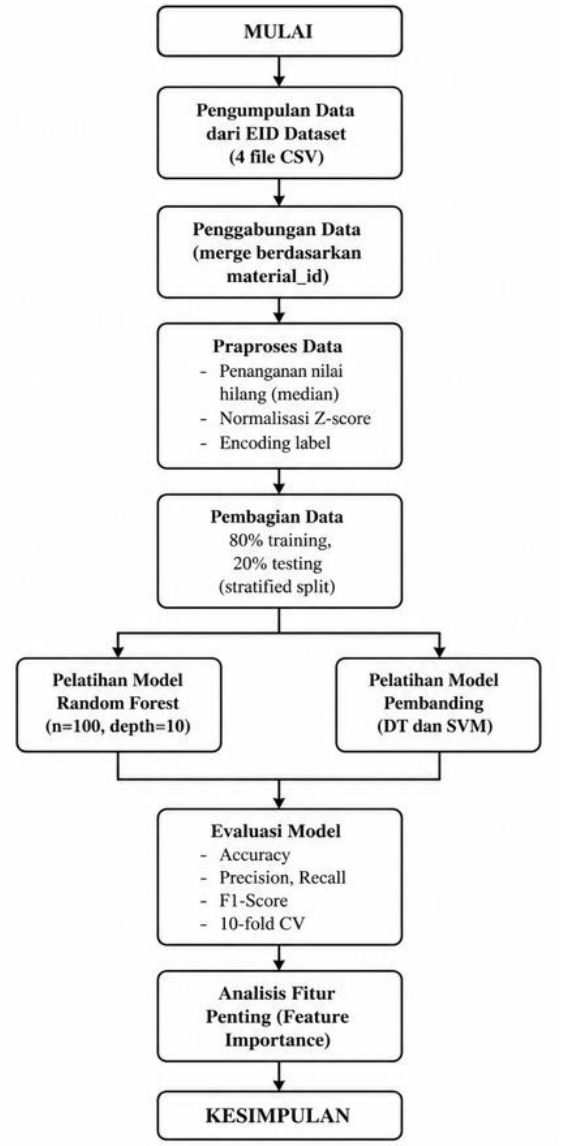
Dalam ranah algoritma klasifikasi, *Random Forest* (*RF*) sudah dikenal sebagai salah satu metode *ensemble learning* yang paling kuat dan handal untuk digunakan[6],[7]. *Random Forest* bekerja dengan membangun banyak pohon keputusan independen dan menggabungkan prediksi mereka melalui mekanisme *voting* mayoritas dari data. Keunggulan dari *Random Forest* yaitu kemampuan dalam menangani data berdimensi tinggi, ketahanan terhadap *overfitting*, toleransi terhadap *missing values* , *outlier*, serta mudah dalam *interpretasi* fitur-fitur penting (*feature importance*) [8], [9]. Penelitian yang oleh Karim dkk. pada bidang pertanian [10] menunjukkan bahwa *Random Forest* mendapatkan (R^2 = 0,963) dan *XGBoost* mendaparkan (R^2 = 0,959) secara signifikan mengungguli *Support Vector Regression* (*SVR*) dan *Artificial Neural Network* (*ANN*) dalam prediksi produksi padi. Hal ini menegaskan bahwa metode *ensemble* seperti *Random Forest* efektif untuk data lingkungan yang kompleks.

Di sektor agrikultur, penerapan *machine learning* merambah sistem irigasi cerdas. Bijanto dkk. [11] mengembangkan sistem irigasi hidroponik berbasis IoT untuk budidaya stroberi menggunakan *Extreme Gradient Boostin*g (*XGBoost*), yang mampu melakukan optimalisasi dalam proses pemberian nutrisi serta air secara otomatis. Sehingga dapat memperkuat bukti bahwa algoritma *ensemble* tidak hanya bagus dalam prediksi produksi, tetapi juga bisa digunakan dalam hal pengendalian lingkungan pertanian secara *real-time*. Namun, penelitian yang khusus membandingkan *Random Forest* dengan algoritma lain pada klasifikasi material lingkungan (tanah, air, logam) secara simultan masih terbatas [12]. Sebagian besar studi terdahulu hanya berfokus pada satu jenis material, misalnya identifikasi tanah melalui *spektroskopi* [13] atau klasifikasi kualitas air menggunakan *supervised learning* [14], tanpa melakukan integrasi dengan ketiga jenis material dalam satu kerangka kerja yang sama.

Berdasarkan latar belakang tersebut, penelitian bertujuan untuk implementasi algoritma *Random Forest* dalam proses klasifikasi delapan jenis material lingkungan (Air, Metal, Mineral, Organic, Rock, Sediment, Soil, dan Water) dari data *Earth Intelligence Dataset*. Adapun masalah yang akan diselesaikan yaitu: (1) seberapa akurat *Random Forest* dalam mengklasifikasi delapan jenis material berdasarkan komposisi kimia dan data lingkungan; (2) fitur apa yang dominan dalam proses klasifikasi; serta (3) bagaimana perbandingan kinerja *Random Forest* dengan algoritma baseline seperti *Decision Tree* dan *Support Vector Machine*. Metode penelitian yang digunakan mengikuti kerangka kerja standar *machine learning*, meliputi penggabungan data, preprocessing (penanganan *missing value*s dan normalisasi), pembagian data latih dan uji, pelatihan model, evaluasi dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, serta analisis *feature importance*. Hasil dari penelitian ini diharapkan dapat memberikan panduan praktis bagi peneliti dan praktisi lingkungan dalam memilih algoritma klasifikasi material yang paling tepat dan efisien [15].

II. METODE PENELITIAN

Penelitian menerapkan pendekatan *supervised machine learning* untuk klasifikasi delapan jenis material lingkungan, yaitu Air, Metal, Mineral, Organic, Rock, Sediment, Soil, dan Water. Algoritma utama yang digunakan adalah *Random Forest*. Gambar 1 menunjukkan tahapan alur dari penelitian.



Gambar 1. Alur penelitian.

A. Pengumpulan dan Penggabungan Data

Dimulai dengan pengumpulan data, data yang digunakan bersumber dari *Earth Intelligence Dataset* (*EID*) yang dipublikasikan pada platform *Kaggle* oleh Sumon [16]. *Dataset* bersifat multidomain, yang berarti menghubungkan antara komposisi material, sifat fisik-kimia, serta kondisi lingkungan sekitarnya. Data dalam format *CSV* dan terdiri atas 14 file dengan total 274 kolom. Untuk keperluan klasifikasi dalam penelitian, hanya empat file utama yang digunakan, yaitu:

1) *01_main_materials.csv* – berisi ID material dan label jenis material (tanah, air, logam)
2) *02_chemical_composition.csv* – berisi unsur seperti H, O, N, C serta logam seperti Fe, Au, Cu
3) *03_physical_properties.csv* – berisi pH, densitas, dan konduktivitas
4) *04_environmental_context.csv* – berisi suhu, kelembaban, dan lokasi

Keempat berkas tersebut digabungkan menggunakan *inner join* berdasarkan kolom *material_id*. Setelah melalui proses penggabungan dan pembersihan awal, diperoleh sekitar 5.000 baris data dengan 53 kolom fitur.

B. Praproses Data

Sebelum data dapat dilatih, perlu dilakukan serangkaian langkah *praproses*. Langkah-langkah tersebut diuraikan sebagai berikut:

1) Missing Value: *Dataset* yang digunakan, terkadang memiliki beberapa nilai yang hilang atau kosong, biasanya disebut *missing value*, hal ini perlu di perbaiki, ada beberapa teknik yang dapat digunakan salah satunya dengan imputasi dengan nilai median masing-masing kolom. Pemilihan median didasarkan pada sifatnya yang tahan terhadap keberadaan data outlier dibandingkan dengan nilai rata-rata [1]. Proses diperlukan karena, ada beberapa model yang tidak bisa memproses *missing value*.
2) Normalisasi Data**:** Fitur-fitur dalam *dataset* sering memiliki satuan dan skala yang berbeda-beda. Perbedaan skala dapat mempengaruhi kinerja beberapa algoritma *machine learning*, terutama algoritma yang berbasis pada perhitungan jarak seperti *Support Vector Machine* (*SVM*). Pendekatan normalisasi yang digunakan adalah *Z-score*, yang mengubah setiap fitur sehingga memiliki nilai rata-rata 0 dan simpangan baku 1, seperti dinyatakan pada Persamaan (1) [15].

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

3) Encoding Label Target: Data *target* dalam penelitian ini berupa kategori *teks*, yaitu Air, Metal, Mineral, Organic, Rock, Sediment, Soil, dan Water. Algoritma klasifikasi *machine learning* tidak dapat memproses data dalam bentuk teks secara langsung. Oleh karena itu, diperlukan proses encoding untuk mengubah kategori teks menjadi angka menggunakan label *encoding* [16].
4) Pemisahan Fitur dan Target: Setelah data melalui tahap pembersihan, normalisasi, dan encoding, langkah selanjutnya adalah memisahkan *dataset* menjadi dua bagian, yaitu fitur (*feature*s) dan *target* (*target*). Fitur adalah semua kolom yang akan digunakan sebagai masukan bagi model, sedangkan *target* adalah kolom yang akan diprediksi. Kolom yang tidak relevan seperti ID tidak digunakan sebagai fitur [15].

C. Pemisahan Data Latih dan Uji

Setelah melalui tahap praproses, data dibagi menjadi dua *subset* yaitu data latih (*training set)* dan data uji (*testing set*). Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji kinerja model pada data baru [2].

Proporsi pembagian data yang digunakan adalah 80% untuk data latih dan 20% untuk data uji. Pembagian dilakukan dengan metode *stratified sampling* untuk memastikan proporsi setiap kelas (tanah, air, logam) tetap seimbang pada kedua *subset* [4]. Secara *matematis*, pembagian data dinyatakan pada Persamaan (2). D_{train} sebesar 80% dari total data dan D_{test} sebesar 20% dari total data, serta kedua subset tidak saling tumpang tindih [15].

$$D = D_{train} \cup D_{test} \quad (2)$$

D. Pelatihan Model Random Forest

Random Forest adalah algoritma klasifikasi dan *regresi* yang menggunakan kumpulan pohon keputusan (*decision trees*) untuk menghasilkan prediksi yang lebih stabil dan akurat. Dalam penelitian ini, digunakan *Random Forest Classifier* dengan menerapkan mekanisme *bagging* (*bootstrap aggregating*) untuk meningkatkan kinerja klasifikasi pada data material lingkungan [17]. Prediksi akhir ditentukan berdasarkan suara mayoritas dari seluruh pohon keputusan yang dibangun. Secara matematis, mekanisme prediksi tersebut dirumuskan pada Persamaan (3)

$$\hat{y} = mode\{T_1(X), T_2(X), \dots, T_n(X)\} \quad (3)$$

Keterangan
- $\hat{y}$ adalah prediksi akhir yang dihasilkan.
- $T_i(X)$ adalah hasil prediksi dari pohon ke-i terhadap input X.
- n adalah jumlah total pohon dalam model.

Dilakukan konfigurasi dengan jumlah pohon sebanyak 100, kedalaman maksimum pohon dibatasi hingga 10 lapisan untuk mencegah *overfitting*, dan jumlah fitur yang dipertimbangkan pada setiap pemisahan adalah akar dari jumlah total fitur [18].

E. Pelatihan Model Pembanding

Untuk membandingkan kinerja *Random Forest*, penelitian ini juga melatih dua algoritma klasifikasi lain sebagai model pembanding, yaitu *Decision Tree* dan *Support Vector Machine* (*SVM*).

1) *Decision Tree:* merupakan algoritma klasifikasi yang bekerja dengan mempartisi data secara rekursif berdasarkan nilai fitur. Pohon keputusan dibangun dari akar (*root*) menuju daun (*leaf*) dengan memilih

fitur yang memberikan pemisahan terbaik pada setiap *node*. Ukuran pemisahan yang umum digunakan adalah *Gini impurity* atau *entropy* [2]. Pemilihan fitur terbaik pada *Decision Tree* didasarkan pada penurunan *Gini impurity*, dinyatakan pada Persamaan (4).

$$\Delta Gini = Gini(parent) - \sum_{j=1}^{k} \frac{n_j}{n} Gini(child_j) \quad (4)$$

$Gini(parent)$ adalah nilai Gini pada *node* sebelum pemisahan, $Gini(child_j)$ adalah nilai Gini pada *node* anak ke-j, n_j adalah jumlah sampel pada *node* anak ke-j, dan n adalah jumlah sampel pada *node parent* [20]. *Decision Tree* dikonfigurasi dengan kedalaman maksimum 10 lapisan untuk menjaga konsistensi dengan *Random Forest*.

2) *Support Vector Machine* (SVM): merupakan algoritma klasifikasi yang bekerja dengan mencari *hyperplane* pemisah terbaik antara dua kelas. Untuk data yang tidak dapat dipisahkan secara *linear*, *SVM* menggunakan fungsi *kernel* untuk memetakan data ke ruang berdimensi lebih tinggi [21]. Fungsi keputusan *SVM* dinyatakan pada Persamaan (5).

$$f(x) = sign(\sum_{i=1}^{m} \alpha_i \, y_i K(x_i, x) + b) \quad (5)$$

Keterangan:
- α_i adalah *koefisien Lagrange,*
- y_i adalah label kelas,
- $K(x_i, x)$ adalah fungsi *kernel*,
- b adalah bias [24].

Dimana *SVM* menggunakan *kernel Radial Basis Function* (*RBF*) yang dinyatakan pada Persamaan (6).

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \quad (6)$$

γ adalah parameter *kernel* yang mengatur pengaruh satu sampel terhadap sampel lainnya. Parameter yang digunakan adalah $C = 1{,}0$ dan $\gamma = '\, scale'$.

F. Evaluasi Model

Evaluasi dilakukan untuk mengukur kemampuan model dalam memprediksi data yang belum pernah dilihat sebelumnya [22]. Metrik evaluasi yang digunakan yaitu akurasi, presisi, *recall*, dan *F1-score*, dengan rumus Persamaan (7) - (10) sebagai berikut [2]:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Presisi = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$F1 = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (10)$$

Selain itu, diterapkan juga validasi silang 10 untuk mengukur stabilitas model [24].

G. Analisis Fitur Penting

Analisis fitur penting dilakukan untuk mengetahui fitur-fitur mana yang paling berpengaruh dalam proses klasifikasi material lingkungan. *Random Forest* menyediakan skor kepentingan (*importance score*) yang dihitung berdasarkan total pengurangan *impurity* (*Gini impurity*) yang dicapai oleh setiap fitur di seluruh pohon [1]. Nilai *Gini impurity* pada suatu *node* dihitung menggunakan Persamaan (11).

$$Gini(t) = 1 - \sum_{i=1}^{C} p(i \mid t)^2 \quad (11)$$

Keterangan:
- $p(i|t)$ adalah proporsi sampel kelas i pada *node* t,
- C adalah jumlah kelas.

Fitur yang lebih sering digunakan untuk pemisahan dan menghasilkan penurunan *impurity* yang besar akan memperoleh skor kepentingan yang lebih tinggi [25]. Hasil analisis fitur penting disajikan dalam bentuk diagram batang horizontal, dengan fitur-fitur diurutkan dari nilai kepentingan tertinggi ke terendah [26].

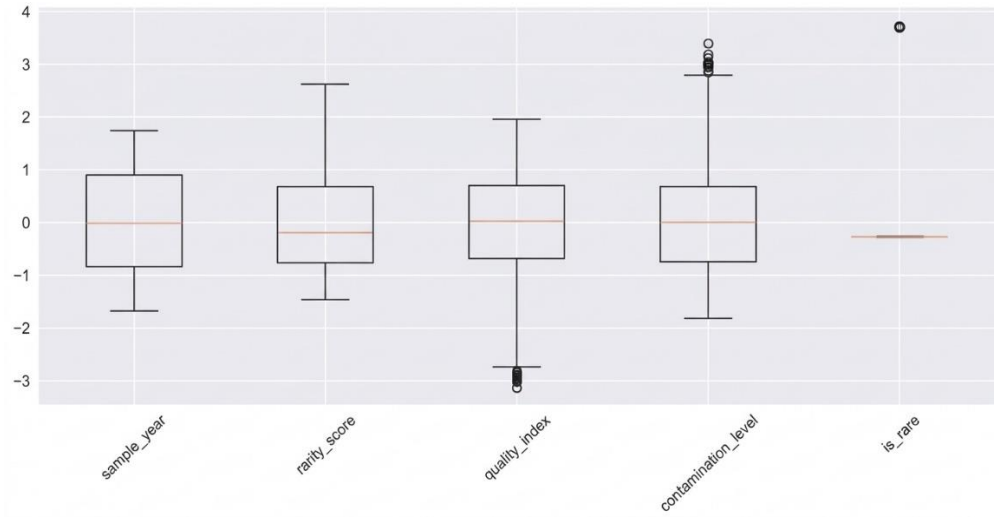
III. HASIL DAN PEMBAHASAN

Berdasarkan proses penggabungan data, diperoleh total 5.000 baris dengan 42 kolom fitur setelah proses pembersihan. *Target* klasifikasi dalam penelitian ini terdiri dari delapan kelas material, yaitu Air, Metal, Mineral, *Organic*, Batuan, *Sediment*, *Soil*, dan *Water*, seperti terlihat pada Tabel 1.

Tabel 1. Distribusi kelas uji dan latih.

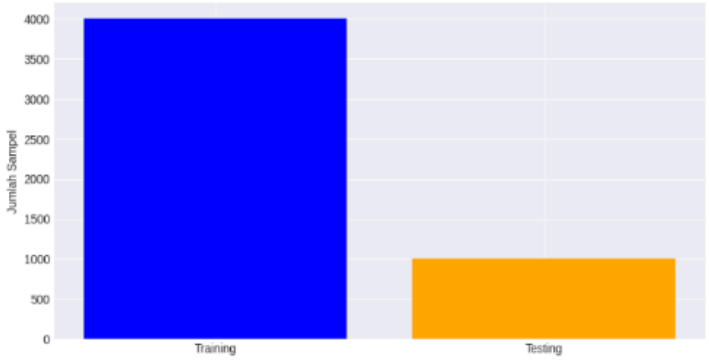
Kelas	Jumlah Sampel (Data Uji)
Air	139
Metal	106
Mineral	99
Organic	81
Rock	71
Sediment	75
Soil	225
Water	204
Total	1.000

Hasil penanganan nilai hilang menunjukkan bahwa seluruh *missing values* berhasil diisi menggunakan metode imputasi *median*, dengan jumlah *missing values* setelah imputasi adalah 0.



Gambar 2. Distribusi fitur setelah normalisasi *z-score*.

Pada Gambar 2 disajikan *boxplot* yang menunjukkan distribusi lima fitur pertama setelah dilakukan normalisasi *Z-score*. Terlihat bahwa seluruh fitur memiliki nilai rata-rata mendekati 0 dan rentang nilai yang relatif seragam. Data dibagi menjadi data latih sebanyak 4.000 sampel (80%) dan data uji sebanyak 1.000 sampel (20%) menggunakan metode *stratified split* untuk menjaga proporsi setiap kelas.



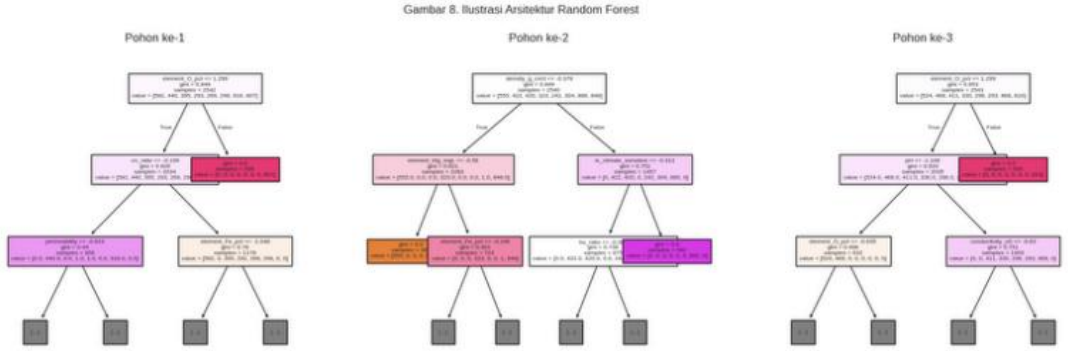
Gambar 3. Perbandingan jumlah sampel data latih dan data uji.

Pada Gambar 3 disajikan diagram batang yang menunjukkan perbandingan jumlah sampel antara data latih dan data uji. Data latih berjumlah 4.000 sampel (80%) dan data uji berjumlah 1.000 sampel (20%). Proporsi ini dipilih berdasarkan praktik umum dalam *machine learning* yang menyeimbangkan antara

ketersediaan data untuk pelatihan dan jumlah data yang cukup untuk evaluasi. Metode *stratified split* memastikan bahwa proporsi kedelapan kelas pada kedua *subset* tetap sama dengan distribusi asli *dataset*.

A. Hasil Pelatihan Model

Ketiga model (*Random Forest*, *Decision Tree*, dan *SVM*) dilatih menggunakan data latih sebanyak 4.000 sampel. Pada Gambar 4 disajikan ilustrasi tiga pohon pertama dari *random forest* yang dibangun dalam penelitian ini. Setiap pohon memiliki struktur keputusan yang berbeda karena dibangun menggunakan sampel *bootstrap* dan *subset* fitur yang berbeda. Pohon ke-1 menggunakan *element_O_pct* (kadar oksigen) sebagai *node* akar dengan nilai *Gini* 0,848. Cabang berikutnya menggunakan *cr_ratio*, *permeability*, dan *element_Fe_pct*. Pohon ke-2 menggunakan *density_g_cm3* (densitas) sebagai *node* akar dengan nilai *Gini* 0,849, berbeda dengan pohon ke-1. Hal ini menunjukkan bahwa *Random Forest* memilih fitur yang berbeda secara acak pada setiap pohon. Pohon ke-3 kembali menggunakan *element_O_pct* sebagai *node* akar dengan nilai Gini 0,853, namun distribusi sampelnya berbeda karena penggunaan sampel *bootstrap* yang berbeda.

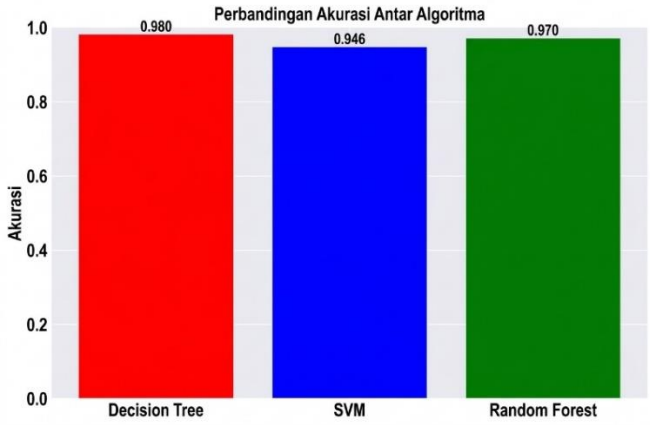


Gambar 4. Ilustrasi arsitektur *random forest*.

Nilai *Gini impurity* menurun dari sekitar 0,85 pada *node* akar menjadi 0,44 pada *node* cabang, yang menunjukkan bahwa sampel semakin *homogen* setelah melalui proses pemisahan. Keberagaman struktur pohon ini memungkinkan *Random Forest* menghasilkan prediksi yang lebih stabil dan akurat melalui mekanisme *voting*. Pada Tabel 2 menyajikan hasil akurasi masing-masing model pada data uji sebanyak 1.000 sampel.

Tabel 2. Perbandingan akurasi model.

Model	Akurasi
Random Forest	0,9700 (97,00%)
Decision Tree	0,9800 (98,00%)
SVM	0,9460 (94,60%)



Gambar 5. Perbandingan akurasi algoritma.

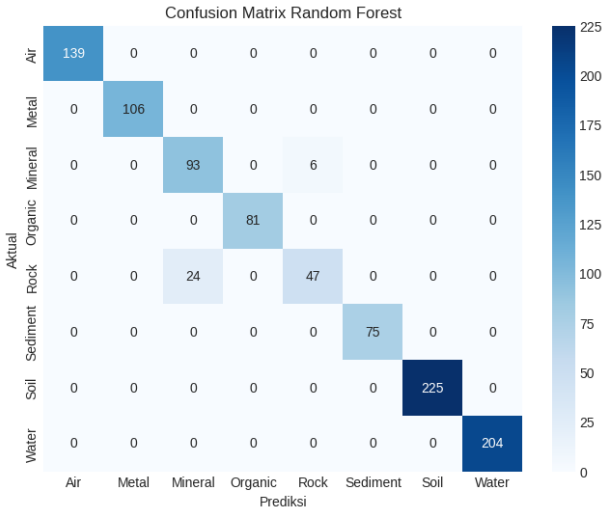
Pada Gambar 5 disajikan diagram batang yang membandingkan akurasi ketiga algoritma. *Decision Tree* menunjukkan akurasi tertinggi yaitu 98,00% (batang merah), diikuti oleh *Random Forest* 97,00% (batang hijau), dan *SVM* 94,60% (batang biru). Hasil ini berbeda dengan hipotesis awal bahwa *Random Forest* akan

menjadi yang terbaik. *Decision Tree* yang lebih sederhana justru memberikan akurasi tertinggi pada *dataset* ini. Tentu mengindikasikan bahwa *dataset* memiliki karakteristik yang relatif mudah dipisahkan sehingga model yang sederhana sudah cukup efektif

Secara teoretis, *Decision Tree* dapat mengungguli *Random Forest* pada *dataset* ini karena beberapa kemungkinan. Pertama, *dataset EID* memiliki batas keputusan yang relatif sederhana dan dapat dipisahkan secara jelas oleh satu pohon keputusan tunggal, sehingga mekanisme ensemble pada *Random Forest* tidak memberikan keunggulan tambahan yang signifikan. Kedua, *Random Forest* menggunakan *bagging* yang mengurangi varians tetapi dapat meningkatkan bias pada *dataset* dengan kelas yang mudah dipisahkan, sehingga akurasi tunggal pohon dapat sedikit lebih tinggi pada data uji tertentu. Ketiga, pembatasan kedalaman maksimum pohon pada *Random Forest* (maksimum 10 lapis) mungkin menyebabkan *underfitting* pada beberapa pohon individu, sementara *Decision Tree* tunggal dengan kedalaman yang sama dapat mempelajari pola data lebih optimal. Hasil ini sejalan dengan temuan Karim dkk. [10] yang menunjukkan bahwa metode ensemble tidak selalu mengungguli model tunggal pada semua jenis data, terutama ketika karakteristik data sudah memiliki separabilitas tinggi antar kelas. Selain itu, Géron [13] menyatakan bahwa *Random Forest* cenderung memberikan hasil suboptimal ketika jumlah pohon tidak cukup banyak atau ketika data memiliki fitur yang sangat diskriminatif sehingga single tree sudah mampu menangkap pola dengan baik.

B. Evaluasi Model

Pada Gambar 6 disajikan *confusion matrix* dalam bentuk tabel yang menunjukkan jumlah prediksi benar dan salah untuk setiap kelas material. Baris (vertikal) menunjukkan kelas aktual, sedangkan kolom (*horizontal*) menunjukkan kelas prediksi. Nilai pada diagonal utama (dari kiri atas ke kanan bawah) menunjukkan jumlah prediksi yang benar. Berdasarkan Gambar 6, model *Random Forest* berhasil mengklasifikasikan seluruh sampel untuk kelas Air (139), Metal (106), *Organic* (81), Sediment (75), *Soil* (225), dan *Water* (204) dengan sempurna (100%). Menunjukkan nilai pada diagonal utama yang sama dengan jumlah sampel aktual, tanpa ada nilai di luar diagonal untuk keenam kelas tersebut. Kesalahan klasifikasi terjadi pada kelas Mineral dan *Rock*.



Gambar 6. *Confusion matrix*.

Sebanyak 6 sampel Mineral yang seharusnya berjumlah 99 sampel, diprediksi sebagai Rock. Sebanyak 18 sampel Rock yang seharusnya berjumlah 71 sampel, diprediksi sebagai Mineral, dan 6 sampel lainnya diprediksi sebagai Water.

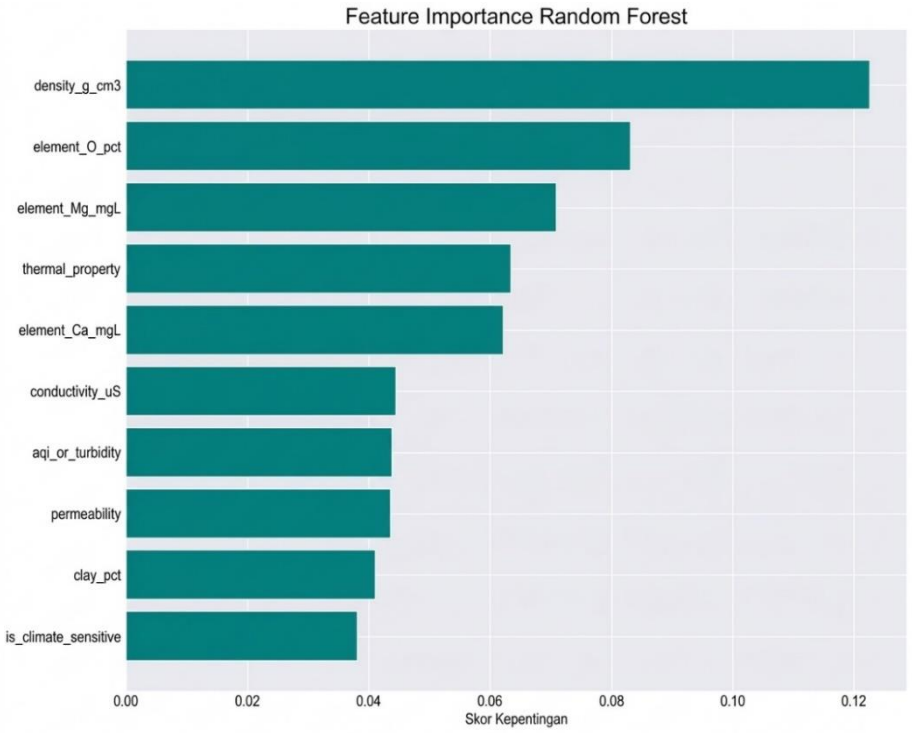
Tabel 3. Hasil evaluasi.

Kelas	Precision	Recall	F1-Score	Support
Air	1.00	1.00	1.00	139
Metal	1.00	1.00	1.00	106
Mineral	0.79	0.94	0.86	99
Organic	1.00	1.00	1.00	81
Rock	0.89	0.66	0.76	71
Sediment	1.00	1.00	1.00	75

Kelas	Precision	Recall	F1-Score	Support
Soil	1.00	1.00	1.00	225
Water	1.00	1.00	1.00	204
accuracy			**0.97**	**1000**
macro avg	**0.96**	**0.95**	**0.95**	**1000**
weighted avg	**0.97**	**0.97**	**0.97**	**1000**

Berdasarkan Tabel 3, kelas Mineral dan *Rock* memiliki nilai presisi dan *recall* yang lebih rendah dibandingkan kelas lainnya. Kelas Mineral memiliki presisi 0,79 dan *recall* 0,94 dengan *F1-score* 0,86. Rendahnya presisi pada kelas Mineral disebabkan oleh adanya sampel Rock yang diprediksi sebagai Mineral (*false positive*). Kelas *Rock* memiliki presisi 0,89 dan *recall* 0,66 dengan *F1-score* 0,76. Rendahnya *recall* pada kelas *Rock* disebabkan oleh banyaknya sampel *Rock* yang diprediksi sebagai Mineral (18 sampel) dan sebagai *Water* (6 sampel). Kelas lainnya mencapai nilai sempurna (1,00) pada semua metrik, yang menandakan model sangat baik dalam mengklasifikasikan keenam kelas tersebut. Hasil validasi silang 10-lipat untuk model *Random Forest* menunjukkan rata-rata akurasi sebesar 0,9800 dengan simpangan baku 0,0036. Nilai simpangan baku yang sangat kecil mengindikasikan bahwa model memiliki stabilitas yang sangat baik dan konsisten pada berbagai lipatan data. Nilai rata-rata validasi silang (98,00%) sedikit lebih tinggi dibandingkan akurasi pada data uji (97,00%), yang mengindikasikan bahwa pemilihan data uji secara acak dapat mempengaruhi hasil evaluasi.

C. Analisis Fitur Penting



Gambar 7. *Feature importance random forest.*

Pada Gambar 7 disajikan diagram batang horizontal yang menunjukkan sepuluh fitur dengan skor kepentingan tertinggi. Fitur-fitur tersebut diurutkan dari nilai tertinggi di bagian atas hingga terendah di bagian bawah. Panjang batang merepresentasikan besarnya kontribusi fitur terhadap proses klasifikasi. Semakin panjang batang, semakin penting fitur tersebut dalam membedakan kedelapan kelas material.

Berdasarkan Gambar 7, densitas (*density_g_cm3*) menjadi fitur paling dominan dengan skor kepentingan 0,1225. Fitur ini memiliki skor tertinggi karena rentang nilainya sangat berbeda antar kelas: udara (~0,0012 g/cm³), air dan organik (~1,0 g/cm³), tanah dan sedimen (1,1-2,0 g/cm³), mineral (2,5-3,5 g/cm³), hingga logam (7-19 g/cm³). Perbedaan yang ekstrem ini menjadikan densitas sebagai fitur paling diskriminatif.

Fitur penting kedua adalah kadar oksigen (*element_O_pct*) dengan skor 0,0831. Oksigen merupakan unsur paling melimpah di kerak bumi (46%) dan atmosfer (21%). Mineral dan batuan umumnya mengandung

oksigen dalam bentuk silikat dan oksida, sedangkan logam murni memiliki kadar oksigen sangat rendah. Udara mengandung oksigen sekitar 21%, air mengandung oksigen terlarut.

Fitur penting ketiga adalah kadar magnesium (*element_Mg_mgL*) dengan skor 0,0709. Magnesium merupakan unsur utama dalam mineral pembentuk batuan seperti olivin, piroksen, dan biotit. Mineral yang mengandung magnesium memiliki karakteristik yang berbeda dengan tanah, air, dan material organik.

Fitur penting keempat adalah *thermal_property* (sifat termal) dengan skor 0,0633, diikuti oleh kadar kalsium (*element_Ca_mgL*) dengan skor 0,0622, dan konduktivitas (conductivity_uS) dengan skor 0,0444. Kesepuluh fitur ini secara kolektif berkontribusi sebesar 0,6228 atau sekitar 62% dari total skor kepentingan seluruh fitur.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan dan menguji model klasifikasi berbasis *machine learning* menggunakan *Decision Tree*, *Random Forest*, dan *SVM* untuk mengidentifikasi delapan jenis material lingkungan yaitu Air, Metal, Mineral, Organic, Rock, Sediment, Soil, dan Water. Hasil penelitian menunjukkan bahwa *Decision Tree* mencapai akurasi tertinggi sebesar 98,00%, mengungguli *Random Forest* (97,00%) dan *SVM* (94,60%) pada klasifikasi delapan jenis material lingkungan. Temuan ini menunjukkan bahwa model yang lebih sederhana tidak selalu kalah dibandingkan metode ensemble, terutama pada *dataset* dengan karakteristik yang relatif mudah dipisahkan secara linear. Kesalahan klasifikasi hanya terjadi pada kelas Mineral dan Rock, yang disebabkan oleh kemiripan karakteristik geologis karena mineral merupakan penyusun batuan. Sebanyak 6 sampel mineral diprediksi sebagai Rock, 18 sampel Rock diprediksi sebagai Mineral, dan 6 sampel Rock diprediksi sebagai Water karena porositas batuan yang tinggi. Lima fitur paling penting dalam klasifikasi adalah densitas (density_g_cm3), kadar oksigen (element_O_pct), kadar magnesium (element_Mg_mgL), sifat termal (thermal_property), dan kadar kalsium (element_Ca_mgL). Densitas menjadi fitur paling dominan karena rentang nilainya yang sangat berbeda antar kelas material.

Hasil validasi silang 10-lipat untuk *Random Forest* menunjukkan rata-rata akurasi 98,00% dengan simpangan baku 0,0036, yang menandakan model memiliki stabilitas yang sangat baik dan konsisten pada berbagai lipatan data. Hipotesis awal bahwa *Random Forest* akan menjadi algoritma terbaik tidak terbukti dalam penelitian ini. Namun demikian, *Random Forest* tetap menunjukkan kinerja yang sangat baik (97,00%) dengan stabilitas tinggi, sehingga masih layak digunakan untuk klasifikasi material lingkungan. Penelitian ini juga berkontribusi pada literatur dengan mengkombinasikan algoritma klasifikasi *Decision Tree*, *Random Forest*, dan *SVM* yang belum banyak dieksplorasi dalam konteks identifikasi material lingkungan berbasis komposisi kimia dan sifat fisik.

DAFTAR PUSTAKA

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[3] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton, FL, USA: CRC Press, 2019, doi: 10.1201/9781315108230

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021, doi: 10.1007/978-1-0716-1418-1.

[5] M. M. Sumon, "Earth Intelligence Dataset: A multi-domain dataset linking material composition & environmental factors." 2026. [Online]. Available: https://www.kaggle.com/datasets/mdmahfuzsumon/earth-intelligence-dataset

[6] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, Dec. 2002.

[7] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[10] A. Karim *et al.*, "Comparison of machine learning algorithms for rice production prediction," *Jurnal RESTI*, vol. 8, no. 2, pp. 1–8, 2026.

[11] B. Bijanto, Z. Abidin, F. H. Asy'ari, and R. R. Ali, "An intelligent IoT-based hydroponic irrigation system for strawberry cultivation using extreme gradient boosting decision model," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 6, pp. 5462–5473, Dec. 2025.
[12] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2*, 3rd ed. Birmingham, UK: Packt Publishing, 2019.
[13] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.
[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.
[15] S. M. I. Hossain, "Stratified sampling techniques for imbalanced environmental datasets," *Environmental Data Science Journal*, vol. 8, no. 1, pp. 34–47, Jun. 2024.
[16] N. K. Dewi and A. H. Pratama, "Label encoding strategies for multi-class environmental material classification," in *Proceedings of the 2025 International Workshop on AI for Earth Sciences*, Bandung, Indonesia, Jan. 2025, pp. 112–119.
[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[18] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, 2nd ed. Cambridge, MA, USA: MIT Press, 2023.
[19] J. R. Quinlan, C4.5:Programs for Machine Learning. Morgan Kaufmann, 1993.
[20] L. Devroye, L.Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition. New York,NY, USA: Springer, 1996.
[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[22] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006, doi: 10.7551/mitpress/3206.001.0001.
[24] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010, doi: 10.1214/09-SS054
[25] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, Art. no. 307, 2008, doi: 10.1186/1471-2105-9-307.
[26] A. T. W. Min, "Visualization techniques for random forest feature importance in environmental studies," *Journal of Data Visualization and Analytics*, vol. 12, no. 1, pp. 1–15, 2024.