Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

Prediksi Periode Fosil Trilobita Menggunakan XGBoost dengan Seleksi Fitur Geologi–Geospasial dan Hyperparameter Tuning

Naufal Rizky Ramadhan, Elkaf Rahmawan Pramudya*

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Kota Semarang, Indonesia
Email: [1]111202214492@mhs.dinus.ac.id, [2,*]elkaf.rahmawan@dsn.dinus.ac.id
Email Penulis Korespondensi: elkaf.rahmawan@dsn.dinus.ac.id
Submitted: **06/12/2025**; Accepted**: 05/03/2026**; Published: **05/03/2026**

Abstrak−Penelitian ini membahas penerapan algoritma Extreme Gradient Boosting (XGBoost) untuk memprediksi periode umur fosil trilobita berdasarkan data geologi dan geospasial. Permasalahan yang dihadapi dalam penelitian ini adalah tingginya kompleksitas data paleontologi, adanya nilai hilang, serta ketidakseimbangan distribusi kelas pada variabel target *time_period* yang dapat memengaruhi kinerja model prediksi. Tujuan penelitian ini adalah membangun model prediksi periode fosil yang akurat dan stabil melalui tahapan preprocessing data, seleksi fitur, serta optimasi model. Dataset yang digunakan berasal dari Kaggle dan memuat atribut *longitude, latitude, lithology, environment,* dan *collection_type* sebagai fitur utama. Tahapan penelitian meliputi pembersihan data, imputasi nilai hilang, encoding fitur kategorikal, pembagian data menggunakan *stratified train-test split*, serta penanganan ketidakseimbangan kelas melalui pendekatan class weight adjustment. Model XGBoost kemudian dilatih pada data pelatihan dan dioptimalkan menggunakan RandomizedSearchCV untuk memperoleh kombinasi hyperparameter terbaik. Hasil evaluasi pada data pengujian menunjukkan bahwa model XGBoost yang telah dituning mampu mencapai akurasi sebesar 95%, precision sebesar 90%, recall sebesar 93%, dan F1-score sebesar 91%, lebih baik dibandingkan model tanpa tuning. Hasil ini menunjukkan bahwa kombinasi seleksi fitur geologi–geospasial dan hyperparameter tuning pada XGBoost efektif dalam meningkatkan performa prediksi periode fosil trilobita. Hasil penelitian ini diharapkan dapat menjadi pendekatan komputasional pendukung dalam paleontologi untuk membantu penentuan periode fosil secara lebih objektif, efisien, dan terukur berbasis data digital.

Kata Kunci: XGBoost; RandomizedSearchCV; Ketidakseimbangan Data; Trilobita; Paleontologi Digital; Geospasial

Abstract−This study investigates the application of the Extreme Gradient Boosting (XGBoost) algorithm to predict the age period of trilobite fossils based on geological and geospatial data. The challenges addressed in this research include the high complexity of paleontological data, the presence of missing values, and class imbalance in the target variable time_period, which can negatively affect predictive performance. The objective of this study is to develop an accurate and robust fossil age prediction model through systematic data preprocessing, feature selection, and model optimization. The dataset used in this research was obtained from Kaggle and consists of the attributes longitude, latitude, lithology, environment, and collection_type as the main features. The research workflow includes data cleaning, missing value imputation, categorical feature encoding, data splitting using stratified train–test split, and class imbalance handling through a class weight adjustment approach. The XGBoost model was trained on the training dataset and further optimized using RandomizedSearchCV to obtain the optimal hyperparameter configuration. Evaluation results on the testing dataset show that the tuned XGBoost model achieved an accuracy of 95%, precision of 90%, recall of 93%, and an F1-score of 91%, outperforming the model without hyperparameter tuning. These results demonstrate that the integration of geological–geospatial feature selection and hyperparameter tuning in XGBoost is effective in improving the performance of trilobite fossil age period prediction. The results of this study are expected to serve as a computational support approach in paleontology to assist fossil period determination in a more objective, efficient, and data-driven manner.
.

Keywords: XGBoost; RandomizedSearchCV; Imbalance Data; Trilobite; Digital Paleontology; Geospasial

1. PENDAHULUAN

Fosil merupakan bukti fisik yang sangat penting dalam memahami sejarah evolusi makhluk hidup dan perubahan lingkungan bumi dari masa ke masa. Penelitian mengenai prediksi periode fosil memiliki peran strategis dalam ilmu paleontologi dan geologi karena mampu membantu merekonstruksi sejarah evolusi kehidupan dan kondisi paleo lingkungan di masa lalu. Trilobit, sebagai kelompok fosil arthropoda laut yang sangat melimpah dari era Paleozoikum, menjadi biomarker penting untuk penentuan umur lapisan batuan dan studi stratigrafi [1]. Namun, klasifikasi periode fosil trilobit secara presisi masih dihadapkan pada tantangan yang signifikan, terutama karena kompleksitas hubungan antara faktor geologi dan lingkungan pengendapan yang bersifat spasial dan temporal. Ketidaktepatan klasifikasi dapat berdampak signifikan pada pemahaman evolusi, rekonstruksi lingkungan purba, serta aplikasi komersial seperti eksplorasi sumber daya mineral dan energi.

Hingga saat ini proses penentuan periode geologis fosil masih banyak mengandalkan interpretasi manual berdasarkan karakter morfologi dan korelasinya dengan kolom stratigrafi [2]. Pendekatan konvensional yang mengandalkan pengamatan manual kurang efektif untuk mengolah data besar dan multidimensi dari dataset fosil [3]. Tantangan tersebut menuntut adanya metode alternatif berbasis data yang mampu meningkatkan objektivitas dan efisiensi dalam klasifikasi periode fosil.

Metode konvensional yang umum digunakan untuk prediksi umur fosil diantaranya adalah korelasi stratigrafi manual, analisis biostratigrafi berbasis index fossil, radiometric dating, serta identification based on morphological keys yang masih dominan dalam praktik paleontologi [4]. Meskipun teknik tersebut telah mendukung kemajuan sejarah geologi, mereka sangat bergantung pada keahlian individu, memiliki keterbatasan pada data yang tidak

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

lengkap, serta relatif rendah dalam hal skalabilitas dan objektivitas ketika digunakan untuk data fosil digital modern [5].

Seiring meningkatnya digitalisasi dan volume data, metode berbasis machine learning mulai dikembangkan sebagai pelengkap dan alternatif bagi metode prediksi umur fosil konvensional. Algoritma seperti Extreme Gradient Boosting atau biasa dikenal XGBoost telah terbukti mampu menangani data besar, beragam, dan kompleks [6]. Beberapa penelitian telah menerapkan teknik ML dalam domain geosains untuk menyelesaikan permasalahan prediksi stratigrafi, pemetaan geologi, serta geochronological modeling [7]. Namun, fokus penelitian tersebut lebih banyak diarahkan pada prediksi litologi atau karakteristik lingkungan geologi, bukan pada klasifikasi umur geologi fosil. Selain itu, penelitian paleontologi berbasis ML umumnya hanya memanfaatkan fitur taksonomi atau citra fosil sebagai prediktor utama [8]. Penelitian sebelumnya mengembangkan model XGBoost regresi untuk prediksi properti geologis dan menunjukkan peningkatan akurasi dibandingkan metode statistik konvensional [9]. Terdapat penelitian yang menerapkan algoritma Extreme Gradient Boosting (XGBoost) dalam bidang ilmu kebumian, khususnya untuk prediksi parameter geologi seperti porositas dan permeabilitas batuan. Penelitian yang dilakukan menunjukkan bahwa model XGBoost yang dioptimasi mampu meningkatkan nilai koefisien determinasi (R^2) hingga sekitar 0,15 dibandingkan model dasar, serta menurunkan nilai Mean Absolute Error (MAE) secara signifikan pada data logging geologi. Hasil tersebut membuktikan bahwa XGBoost efektif dalam memodelkan hubungan nonlinier dan kompleks pada data geologi berdimensi tinggi, serta mampu memberikan akurasi prediksi yang lebih baik dibandingkan pendekatan statistik konvensional[10]. Penelitian lain menunjukkan bahwa XGBoost mencapai akurasi hingga 90,10% dalam prediksi properti geologi, lebih unggul dibanding metode statistik klasik seperti Naïve Bayes dan SVM [11]. Penelitian lain menyatakan menyatakan XGBoost cocok untuk high-dimensional data dan mampu memberikan hasil prediksi dengan presisi di atas 93% pada klasifikasi ketebalan formasi geologi [12]

Pada sisi metodologi, performa XGBoost sangat dipengaruhi oleh pemilihan *hyperparameter* yang optimal. Pengaturan *hyperparameter* dengan cara manual sering menghasilkan performa suboptimal dan perlu waktu komputasi yang besar[13] Grid Search menjadi teknik yang banyak digunakan, namun kurang efisien ketika ruang parameter sangat luas. RandomizedSearchCV merupakan solusi alternatif yang mampu melakukan eksplorasi ruang parameter secara acak namun tetap terarah, sehingga lebih efisien dalam menemukan kombinasi optimal dengan biaya komputasi yang lebih rendah [14].

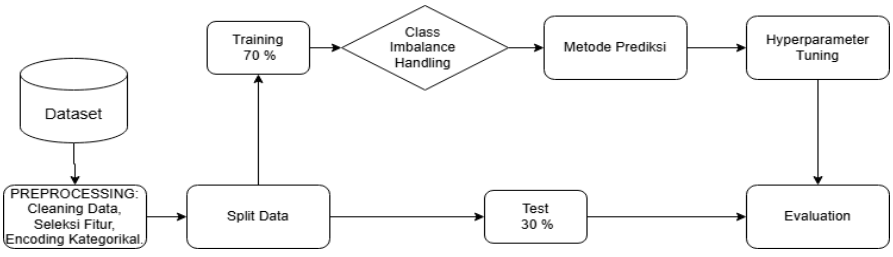
Berdasarkan studi-studi sebelumnya dapat diidentifikasi adanya research gap, yaitu belum adanya penelitian yang mengintegrasikan fitur geologi dan geospasial dalam klasifikasi periode fosil menggunakan XGBoost yang dioptimasi hyperparameter-nya secara terarah. Mayoritas penelitian di bidang paleontologi masih berada pada sisi identifikasi morfologis, sedangkan penelitian pada bidang geoinformatika belum banyak menyentuh permasalahan klasifikasi umur fosil sebagai objek prediksi. Hal ini menunjukkan adanya ruang riset yang masih terbuka luas untuk mengembangkan sistem klasifikasi fosil yang lebih akurat dan objektif melalui pendekatan ML modern.

Oleh karna itu penelitian ini berfokus pada pembuatan model prediksi periode fosil trilobita menggunakan algoritma *Extreme Gradient Boosting* yang dioptimasi menggunakan *RandomizedSearchCV* dengan memanfaatkan integrasi fitur geospasial dan geologi sebagai variabel prediktif utama. Fitur geospasial yang digunakan dalam penelitian ini meliputi koordinat lokasi temuan fosil (longitude dan latitude), sedangkan fitur geologi mencakup jenis litologi, lingkungan pengendapan, dan tipe koleksi. Pendekatan ini diharapkan mampu meningkatkan akurasi klasifikasi dan mengurangi ketergantungan pada interpretasi manual dalam paleontologi. Penelitian ini memiliki kontribusi utama yaitu meningkatkan akurasi prediksi periode geologis pada fosil, lalu mengurangi subjektivitas penentuan periode fosil melalui pendekatan komputasi objektif, dan meyajikan model prediktif yang dapat direproduksi dengan performa yang lebih baik melalui optimasi *hyperparameter*

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan dengan rangkaian tahapan sistematis untuk membangun model prediksi periode fosil trilobita berbasis algoritma *Extreme Gradient Boosting* (XGBoost). Setiap tahapan dirancang agar proses analisis data, pemodelan, hingga evaluasi berjalan dengan terstruktur dan membuahkan model yang akurat pada kondisi data yang memiliki ketidakseimbangan kelas. Gambar 1 menggambarkan keseluruhan proses yang dilakukan, mulai dari pemuatan data sampai tahap evaluasi model



Gambar 1. Diagram Flow

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

Tahapan penelitian diawali dengan proses pemuatan dataset trilobita yang berisi informasi taksonomi, geologi, dan geospasial. Selanjutnya dilakukan tahap pra-pemrosesan awal data yang mencakup pembersihan data, imputasi nilai hilang, encoding fitur kategorikal, serta konversi tipe data agar seluruh atribut dapat diproses oleh model pembelajaran mesin. Setelah proses tersebut, dilakukan seleksi fitur geologi–geospasial dengan memilih fitur *longitude, latitude, lithology, environment,* dan *collection_type* yang dinilai paling relevan terhadap pembentukan periode umur geologi fosil. Dataset yang telah dipra-pemrosesan kemudian dibagi menjadi dua data yaitu pelatihan dan data test menggunakan metode train–test split dengan skema stratified sampling, sehingga proporsi kelas pada target time_period tetap terjaga pada kedua subset data. Penanganan ketidakseimbangan kelas selanjutnya diterapkan hanya pada data pelatihan melalui mekanisme class weight adjustment, bertujuan untuk mengurangi bias model terhadap kelas mayoritas tanpa mengubah distribusi asli pada data pengujian.Model dasar XGBoost kemudian dibangun dan dilatih menggunakan data pelatihan yang telah diberikan bobot kelas. Proses ini dilanjutkan dengan tahap hyperparameter tuning menggunakan RandomizedSearchCV untuk memperoleh kombinasi parameter yang optimal melalui proses cross-validation. Setelah konfigurasi parameter terbaik diperoleh, model kembali dilatih dan dievaluasi menggunakan metrik performa klasifikasi untuk mengukur tingkat akurasi dan stabilitas prediksi periode fosil. Melalui seluruh tahapan penelitian ini, model akhir diharapkan mampu memberikan performa prediksi yang lebih optimal.

2.2 Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset fosil trilobita yang didapat dari situs Kaggle yang berjudul "*Predict the Age of a Trilobite*". Berikut link dari dataset https://www.kaggle.com/datasets/kayleefranklin/predict-the-age-of-a-trilobite/data. Dataset ini berisi kumpulan data fosil trilobit yang telah terdokumentasi dari berbagai negara dan formasi geologi. Secara keseluruhan, dataset ini terdiri dari sekitar 29.039 baris dan 30 kolom atribut yang mencakup informasi mengenai aspek taksonomi, data geospasial, kondisi geologi, serta lingkungan pengendapan fosil tersebut ditemukan. Pada dataset ini terdapat label target yaitu time_period, yang merepresentasikan periode geologis organisme triolobita hidup, seperti *Cambrian, Ordovician, Devonian*, dan periode lainnya. Label ini bersifat multi-kelas dengan jumlah sampel per kelas yang tidak seimbang (*imbalanced*), karena beberapa periode memiliki lebih banyak spesimen dibandingkan periode lain. Ketidakseimbangan kelas ini dapat membuat model menjadi bias terhadap kelas yang menjadi mayoritas sehingga perlu teknik penanganan khusus untuk mengatasi kasus *imbalance* pada dataset ini. Dataset ini digunakan untuk melakukan prediksi pada periode fosil trilobite berdasarkan fitur yang tersedia. Fitur-fitur yang digunakan berfokus pada atribut geologi dan geospasial yang dianggap paling relevan terhadap umur dan lingkungan pembentukan fosil. Dataset ini menjadi dasar penerapan algoritma XGBoost untuk membangun model prediksi yang optimal. Selain itu, dilakukan proses hyperparameter tuning menggunakan RandomizedSearchCV untuk meningkatkan performa model.

2.3 Preprocessing

Tahap preprocessing merupakan proses penting dalam penelitian ini untuk memastikan bahwa data dalam kondisi yang optimal sebelum digunakan dalam pembangunan model. Seluruh tahapan preprocessing dilakukan untuk mengurangi bias model dan memaksimalkan kemampuan generalisasi model pada data uji. Berikut adalah tahapan dari preprocessing

a. Pemuatan Data
   Langkah pertama yaitu pemuatan data, dimana dataset yang berisi data fossil trilobita dimuat kemudian dilakukan pengecekan struktur dataset berupa jumlah baris, kolom, dan tipe data. Analisis awal ini penting untuk memahami kondisi kualitas data sebelum pemrosesan lanjutan [15].
b. Pembersihan Data
   Langkah selanjutnya yaitu pembersihan data awal, dengan melakukan penghapusan baris yang tidak memiliki nilai pada kolom target *time_period*, karena model klasifikasi tidak dapat melatih prediksi tanpa label. Langkah ini penting untuk mendukung peningkatan validitas selama pelatihan [16].
c. Seleksi Fitur Geologi-Geospasial
   Pada tahap ini dilakukan pemilihan atau seleksi fitur yang dianggap memiliki pengaruh paling relevan terhadap penentuan periode fosil trilobita. Fitur yang dipilih meliputi *longitude, latitude, lithology, environment,* dan *collection_type*. Dalam penelitian ini metode seleksi fitur yang digunakan adalah berdasarkan Pengetahuan Domain (*Domain Knowledge*).Fitur-fitur seperti *'Longtitude', 'Latitude', 'Lithology', 'Environment',* dan *'Collection_type*' dipilih karena secara geologis dan paleontologis dianggap sebagai indikator kuat yang mempengaruhi *'time period*' fosil trilobita[17]. Pemilihan ini bertujuan untuk memfokuskan model hanya pada variabel yang berhubungan langsung dengan kondisi geologi dan lokasi pengendapan fosil. Dengan membatasi fitur hanya pada atribut yang relevan, proses pelatihan model menjadi lebih efisien, mengurangi risiko *overfitting*, serta meningkatkan akurasi prediksi periode geologis fosil [4]
d. Encoding Kategorikal
   Pada tahap ini dilakukan proses encoding fitur kategorikal agar dapat dipahami oleh model XGBoost yang hanya menerima input berupa numerik. Fitur kategorikal yang digunakan, yaitu *lithology, environment*, dan *collection_type*, diubah menjadi representasi numerik menggunakan *One-Hot Encoding*. Metode ini dipilih karena

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

tidak memberikan asumsi urutan pada kategori serta efektif digunakan dalam pemodelan data tabular berbasis pohon keputusan. Tujuan utama tahap ini adalah memastikan setiap kategori dapat diproses dengan benar oleh model tanpa menimbulkan bias struktural antar kelas [18].

e. Train-Test Split
Sebelum melanjutkan pada langkah penanganan data yang *imbalance*, pada tahap ini dilakukan pemisahan dataset menjadi data pelatihan dan data pengujian dengan proporsi 70% untuk pelatihan dan 30% untuk pengujian. Teknik *stratified sampling* diterapkan untuk memastikan bahwa distribusi kelas pada variabel target *time_period* tetap konsisten pada kedua subset data. Tujuan dari proses ini adalah menjaga representativitas data sehingga model dapat mempelajari pola dengan baik dan menghindari bias akibat distribusi kelas yang tidak seimbang antara data pelatihan dan pengujian [18].

f. Penanganan Data Imbalance
Setelah melakukan split data, diperlukan penanganan data yang imbalance atau tidak seimbang pada data training, seperti pada data yang digunakan ini mengalami ketidakseimbangan kelas pada variabel target *time_period* yang memiliki distribusi jumlah kelas tidak merata. Seperti pada dataset yang digunakan untuk penelitian ini, periode *Ordovician* merupakan kelas yang paling dominan dengan 43.55% dari total sampel pelatihan, sedangkan periode *Permian* hanya menyumbang 1.51% dan *Carboniferous* sebesar 3.33%. Ketidakseimbangan ini dapat menyebabkan model bias terhadap kelas dengan jumlah sampel terbesar sehingga menurunkan kualitas generalisasi prediksi [19].Oleh karena itu, penelitian ini menggunakan *class weight adjustment* pada algoritma XGBoost, di mana bobot kelas dihitung berdasarkan kebalikan dari proporsi tiap kelas sesuai Persamaan (1)

$$W_{c=}\frac{N}{K.n_C} \tag{1}$$

Dimana N adalah total sampel dan K adalah jumlah kelas. Pendekatan *cost-sensitive* learning ini dipilih karena lebih efektif untuk *multi-class classification*, serta tidak menambah sampel sintetis yang berisiko menimbulkan noise pada fitur kategorikal, seperti yang umumnya terjadi pada metode oversampling berbasis SMOTE [20].

2.4 Extreme Gradient Boost

Extreme Gradient Boost (XGBoost) merupakan algoritma berbasis pohon keputusan (*decision tree ensemble*) yang merupakan pengembangan dari teknik *Gradient Boosting Machine* (GBM), dan XGBoost dapat menangani permasalahan baik klasifikasi maupun regresi [21]. Pada struktur pohon regresi, setiap internal node berfungsi sebagai penguji atribut dengan ambang batas tertentu, sedangkan *leaf node* berperan dalam memberikan skor hasil keputusan akhir dari model [22] . Pada setiap iterasi, algoritma membangun pohon keputusan baru yang berfokus pada sisa kesalahan (*residual error*) dari prediksi sebelumnya. Proses ini terus berlanjut hingga kesalahan model mencapai titik minimum. Dalam regresi, fungsi objektif XGBoost dapat sebagaimana ditunjukkan pada Persamaan (2)

$$Obj = \sum_{i=1}^{n} l(y_i,\ y\hat{}_i) + \sum_{k=1}^{K} \Omega(fk) \tag{2}$$

Dimana fungsi $\sum_{i=1}^{n} l(y_i, y\hat{}_i)$ adalah fungsi loss yang dipergunakan untuk menghitung dan mengukur seberapa baik model untuk menyesuaikan terhadap data latih. Fungsi $\sum_{k}^{K} \Omega(fk)$ adalah fungsi regulasi untuk mengendalikan kompleksitas model, fungsi ini berperan penting untuk mencegah *overfitting*, yaitu kondisi di mana model terlalu menyesuaikan diri pada data pelatihan.

2.5 Hyperparameter Tuning

Tahap tuning *hyperparameter* merupakan proses penting untuk memperoleh kombinasi parameter terbaik yang dapat memaksimalkan performa model *Extreme Gradient Boosting (XGBoost)*[23] . Meskipun algoritma XGBoost telah memiliki kemampuan prediktif yang tinggi secara default, nilai-nilai parameter yang tidak diatur dengan tepat dapat menyebabkan model tidak mencapai potensi optimalnya atau bahkan mengalami overfitting [24]. Oleh karena itu, proses optimasi dilakukan dengan pendekatan sistematis menggunakan metode RandomizedSearchCV. Metode RandomizedSearchCV bekerja dengan cara melakukan pencarian acak (*random sampling*) terhadap kombinasi parameter dari ruang pencarian (*search space*) yang telah ditentukan sebelumnya. Berbeda dengan GridSearchCV yang menguji seluruh kemungkinan kombinasi parameter secara menyeluruh, RandomizedSearchCV memilih sejumlah kombinasi acak yang representatif [25]. Penyesuaian parameter-parameter tersebut bertujuan untuk menyeimbangkan kompleksitas model dan kemampuan generalisasi sehingga model mampu menangkap pola geologi yang relevan tanpa menyesuaikan diri secara berlebihan terhadap data pelatihan. Pendekatan ini terbukti jauh lebih efisien secara komputasi, terutama pada dataset berukuran besar seperti dataset fosil trilobita yang memiliki variasi fitur geospasial dan geologi yang kompleks. Parameter utama yang dilakukan optimasi pada penelitian ini meliputi :

a. n_estimators: Jumlah pohon dalam model. Umumnya berkisar antara 100 hingga 1000 atau lebih, tergantung kompleksitas data dan waktu komputasi yang tersedia. Pada penelitian ini digunakan rentang 100-900 agar model cukup kuat mempelajari pola data tanpa membutuhkan waktu komputasi yang terlalu besar.
b. learning_rate: Tingkat pembelajaran, mengecilkan ukuran langkah di setiap iterasi. Umumnya berkisar antara 0.01 hingga 0.3. Pada penelitian ini digunakan rentang 0.01-0.3 untuk menjaga keseimbangan antara stabilitas pelatihan dan kecepatan konvergensi model.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

c. max_depth: Kedalaman maksimum dari setiap pohon. Umumnya antara 3 hingga 10. Pada penelitian ini digunakan rentang 3-9 untuk mencegah model menjadi terlalu kompleks dan mengurangi risiko overfitting.
d. colsample_bytree: Rasio sub-sampel kolom saat membangun setiap pohon. Umumnya antara 0.6 hingga 1.0. Pada penelitian ini digunakan rentang 0.7-1.0 agar model lebih beragam dan tidak terlalu bergantung pada fitur tertentu.
e. subsample: Rasio sub-sampel instance pelatihan per pohon. Umumnya antara 0.6 hingga 1.0. Pada penelitian ini digunakan rentang 0.7-1.0 untuk membantu mengurangi overfitting dan meningkatkan kemampuan generalisasi model.
f. gamma: Parameter minimal pengurangan loss yang diperlukan untuk membuat partisi lebih lanjut pada node daun pohon. Umumnya antara 0.0 hingga 0.2, seringkali dimulai dari 0. Pada penelitian ini: digunakan rentang 0.0-0.2 untuk mengontrol kompleksitas pohon dan menghindari pemisahan yang kurang penting.

Proses tuning yang dilakukan pada penelitian ini menggunakan teknik validasi silang (*cross-validation*) sebanyak tiga lipatan (k-fold = 3) Penelitian ini menggunakan teknik validasi silang (*cross-validation*) sebanyak tiga lipatan (k-fold = 3) dalam proses tuning hyperparameter dengan RandomizedSearchCV dengan alasan strategis yaitu, teknik ini menjaga keseimbangan optimal antara efisiensi komputasi dan kualitas evaluasi, menghindari waktu proses yang berlebihan yang akan terjadi dengan jumlah lipatan yang lebih tinggi, sementara tetap mengeksplorasi beragam kombinasi hyperparameter secara efektif[26].

Banyak praktisi dan literatur machine learning merekomendasikan k=3, k=5, atau k=10 sebagai nilai default yang baik untuk validasi silang, kecuali ada alasan khusus untuk memilih nilai lain (misalnya, jika dataset sangat kecil, k yang lebih besar mungkin diperlukan untuk memastikan setiap lipatan memiliki sampel yang cukup) [27]. Dalam kasus ini, dengan 26.074 entri, k=3 masih memberikan ukuran fold yang cukup besar (sekitar 8.600 entri per fold untuk pelatihan dan 4.300 entri untuk pengujian di setiap iterasi), yang dianggap representatif. Proses ini menghasilkan kombinasi parameter terbaik yang digunakan kembali dalam pelatihan model akhir. Pendekatan RandomizedSearchCV dipilih karena memberikan keseimbangan antara efisiensi komputasi dan kualitas hasil optimasi [24]. Penelitian terbaru menunjukkan bahwa metode ini mampu menghasilkan hasil yang setara bahkan lebih baik dibandingkan GridSearchCV dengan waktu komputasi yang lebih singkat [14]

2.6 Evaluasi

Pada penelitian ini evaluasi model dilakukan untuk menilai sejauh mana algoritma Extreme Gradient Boosting (XGBoost) mampu memprediksi periode fosil trilobita secara akurat setelah melalui proses hyperparameter tuning menggunakan RandomizedSearchCV. Tahapan evaluasi ini bertujuan untuk mengukur kinerja model berdasarkan data pengujian (*testing set*) yang belum pernah dilihat sebelumnya selama proses pelatihan. Evaluasi kinerja dilakukan menggunakan beberapa metrik utama yang umum digunakan dalam permasalahan klasifikasi multikelas, yaitu akurasi, precision, recall, dan F1-score, yang dihitung berdasarkan confusion matrix. Dalam konteks ini, *True Positive* (TP) merepresentasikan jumlah data fosil yang diprediksi ke dalam suatu periode tertentu dan sesuai dengan kelas sebenarnya, sedangkan *True Negative* (TN) menunjukkan jumlah data yang secara benar diprediksi tidak termasuk ke dalam periode tersebut. Sementara itu, *False Positive* (FP) adalah kondisi ketika data fosil diprediksi ke dalam suatu periode tertentu namun kelas sebenarnya berbeda, dan *False Negative* (FN) merepresentasikan data fosil yang seharusnya termasuk dalam suatu periode tetapi gagal dikenali oleh model. Keempat komponen ini menjadi dasar perhitungan metrik evaluasi dan memberikan gambaran menyeluruh mengenai ketepatan, kelengkapan, dan keseimbangan performa model dalam melakukan prediksi periode fosil trilobita.

a. Akurasi
Akurasi digunakan untuk mengukur proporsi prediksi yang benar dibandingkan dengan seluruh prediksi yang dilakukan oleh model [28]. Perhitungan nilai akurasi pada penelitian ini dilakukan menggunakan Persamaan (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

b. Presisi
Presisi menggambarkan sejauh mana model mampu memberikan prediksi yang tepat pada setiap kelas tanpa banyak menghasilkan kesalahan positif[29]. Presisi dihitung menggunakan Persamaan (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

c. Recall
Recall menunjukkan kemampuan model dalam mengenali seluruh sampel dari kelas sebenarnya, sementara [30]. Nilai recall dihitung berdasarkan Persamaan (5).

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

d. F1-Score
F1-score merupakan rata-rata harmonis antara presisi dan recall yang memberikan gambaran keseimbangan performa model pada setiap kelas[31]. Perhitungan F1-score pada penelitian ini mengacu pada Persamaan (6).

$$F1-Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (6)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

3. HASIL DAN PEMBAHASAN

3.1 Hasil Preprocessing

Pada tahap preprocessing,dilakukan serangkaian prosedur untuk memastikan bahwa dataset yang digunakan berada dalam kondisi optimal sebelum memasuki proses pemodelan. Tahapan preprocessing mencakup pembersihan data, seleksi fitur geologi-geospasial, penanganan data imbalance menggunakan metode class weight adjustment, dan pembagian data untuk training dan testing. Hasil dari setiap tahapan akan dijelaskan secara detail.

3.1.1 Pemuatan Dataset

Pada penelitian ini, dataset yang digunakan adalah dataset fosil trilobita yang diperoleh dari situs Kaggle dengan judul "*Predict the Age of a Trilobite*". Dataset ini memiliki 29.039 baris dan 30 kolom yang mencakup informasi mengenai aspek taksonomi, data geospasial, kondisi geologi, serta lingkungan pengendapan fosil tersebut ditemukan. Pada tahap ini, dataset trilobita dimuat ke dalam lingkungan pemrograman Python menggunakan library pandas. Proses ini juga mencakup pengecekan struktur dataset seperti jumlah baris, jumlah kolom, tipe data, serta pemeriksaan awal terhadap keberadaan nilai hilang Tabel 1 berikut berisi dataset yang digunakan dalam penelitian ini

Tabel 1. Data Fosil Trilobita

No	scientific_name	order	order_num	...	vision	diet	time_period
0	Australosutura llanoensis	Proetida	21062	...	well-developed	deposit feeder	Carboniferous
1	Phillibole planucauda	Proetida	21062	...	well-developed	deposit feeder	Carboniferous
2	Thigriffides roundyi	Proetida	21062	...	well-developed	deposit feeder	Carboniferous
3	Pudoproetus chappelensis	Proetida	21062	...	well-developed	deposit feeder	Carboniferous
4	Pudoproetus chappelensis	Proetida	21062		well-developed	deposit feeder	Carboniferous
...	...	...	...	...	...	...	...
29035	Pseudophillipsia caucasia	Proetida	21062	...	well-developed	deposit feeder	Permian
29036	Pseudophillipsia caucasia	Proetida	21062	...	well-developed	deposit feeder	Permian
29037	Pseudophillipsia caucasia	Proetida	21062	...	well-developed	deposit feeder	Permian
29038	Pseudophilipsia solida	Proetida	21062	...	well-developed	deposit feeder	Permian
29039	Pseudophilipsia solida	Proetida	21062	...	well-developed	deposit feeder	Permian

3.1.2 Pembersihan Data

Setelah data awal dimuat, ditemukan adanya nilai yang hilang (NaN) pada kolom target *time_period*. Karena nilai hilang ini akan menghambat proses pelatihan model dan evaluasi, langkah pembersihan data krusial pertama adalah menghapus baris-baris yang mengandung nilai NaN pada kolom *time_period* untuk memastikan integritas data target sebelum seleksi fitur geologi-geospasial yang lebih spesifik. Langkah ini dilakukan agar setiap sampel yang digunakan dalam pelatihan memiliki label periode geologi yang jelas dan valid. Dengan demikian, model dapat belajar dari data yang konsisten dan terhindar dari bias yang disebabkan oleh ketidaklengkapan informasi target, seperti pada Tabel 2.

Tabel 2. Hasil Pembersihan Data

No	Tahap Data	Jumlah Baris total	Jumlah NaN di time_period
0	Data Asli	29039	2965
1	Data Setelah Pembersihan	26074	0

Tabel 2 menunjukan hasil dari pembersihan data, dimana dataset telah dibersihkan dari data yang hilang atau data NaN yang terdapat pada kolom target yaitu *time_period*. Terungkap terdapat 2965 data yang hilang pada kolom target *time_period*, setelah dilakukan pembersihan data jumlah awal baris total 29039 menjadi 26074.

3.1.3 Seleksi Fitur Geologi-Geospasial

Setelah tahap pembersihan data awal yang fokus pada kolom target, langkah selanjutnya adalah melakukan seleksi fitur dari dataset yang tersedia. Seleksi fitur ini krusial untuk memastikan bahwa model hanya dilatih menggunakan informasi yang paling relevan, sehingga meningkatkan efisiensi komputasi dan potensi akurasi prediksi. Untuk kasus prediksi periode fosil trilobita ini, fitur-fitur yang dipilih secara spesifik mencakup aspek geologi dan geospasial yang diyakini memiliki korelasi kuat dengan distribusi waktu fosil yang ditampikan pada tabel 3.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

Tabel 3. Fitur yang Dipilih

No	Fitur	Deskripsi
1	Longtitude	Merepresentasikan lokasi geografis lintang
2	Latitude	Merepresentasikan lokasi geografis bujur
3	Lithology	Mengacu pada karakteristik fisik batuan tempat fosil ditemukan
4	Environment	:Menjelaskan jenis lingkungan pengendapan
5	Collection_type	Menunjukkan jenis koleksi atau metode pengumpulan data fosil,

Terlihat pada Tabel 3 fitur-fitur itu dipilih karena secara langsung terkait dengan kondisi lingkungan dan geologis di mana trilobite hidup dan terfosilisasi, yang merupakan indikator kuat untuk periode waktunya. Dengan memusatkan perhatian pada fitur-fitur ini, kita dapat membangun model yang lebih fokus dan relevan dengan tujuan prediksi

3.1.4 Split Data

Setelah tahapan pra-pemrosesan awal data, yang meliputi pembersihan nilai hilang dan seleksi fitur, dataset selanjutnya dipisahkan menjadi dua bagian utama, yaitu data pelatihan (training set) dan data pengujian (testing set). Penanganan ketidakseimbangan kelas kemudian diterapkan hanya pada data pelatihan melalui pemberian sample_weight, sehingga distribusi kelas pada data pengujian tetap merepresentasikan kondisi data asli dan digunakan secara objektif untuk evaluasi performa model.Seperti yang terlihat di Tabel 4.

Tabel 4. Pembagian Data Training dan Testing

No	Dataset	Jumlah
0	X_train	(18251,77)
1	y_train_encoded	(18251,)
2	Sample_weight_train	18251
3	X_test	(7823,77)
4	y_test_encoded	(7823,)

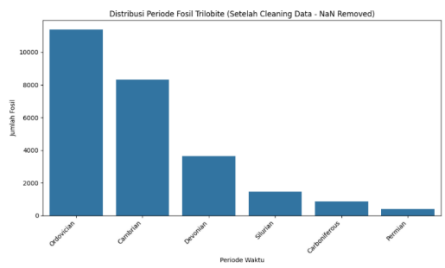
Seperti yang terlihat pada Tabel 4 pembagian ini dilakukan dengan proporsi 70% untuk data pelatihan dan 30% untuk data pengujian. Tujuan utama dari pemisahan ini adalah untuk melatih model pada sebagian data dan mengevaluasinya secara independen pada bagian data yang belum pernah dilihat oleh model, sehingga dapat mengukur kemampuan generalisasi model.

3.1.5 Penanganan Data Imbalance

Setelah seluruh tahapan pra-pemrosesan data, termasuk pembersihan nilai hilang, seleksi fitur,dan pembagian data training dan testing, dilakukan pemeriksaan terhadap distribusi kelas pada variabel target time_period. Beberapa periode geologi memiliki jumlah sampel yang jauh lebih besar dibandingkan periode lainnya, sehingga berpotensi menyebabkan model lebih dominan mempelajari kelas mayoritas. Kondisi ini dapat menurunkan kemampuan model dalam mengenali kelas minoritas jika tidak ditangani dengan tepat. Oleh karena itu, diperlukan strategi khusus agar model tetap mampu memberikan prediksi yang adil dan akurat pada seluruh kelas periode fosil. Hasil analisis menunjukkan adanya ketidakseimbangan yang signifikan antar kelas, seperti yang tertera pada Gambar 2 dan Tabel 5.

Tabel 5. Visualiasasi Penanganan Imbalance Data

No	Time Period	Percentage	Sample Weight
0	Ordovician	43.633505	0.381969
1	Cambrian	31.890005	0.522630
2	Devonian	13.994784	1.190920
3	Silurian	5.630130	2.960263
4	Carboniferous	3.305975	5.041377
5	Permian	1.545601	10.783292



Gambar 2. Data Imbalance

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

Terlihat pada Gambar 2 dan Tabel 5, kelas 'Ordovician' dan 'Cambrian' merupakan kelas mayoritas dengan proporsi yang jauh lebih besar dibandingkan kelas minoritas seperti 'Carboniferous' dan 'Permian'. Ketidakseimbangan ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas dan memiliki kinerja yang buruk dalam memprediksi kelas minoritas. Untuk mengatasi masalah ketidakseimbangan data ini, digunakan strategi pemberian bobot sampel (sample_weight). Bobot sampel dihitung berdasarkan invers proporsi frekuensi setiap kelas. Kelas-kelas minoritas diberi bobot yang lebih tinggi, sementara kelas-kelas mayoritas diberi bobot yang lebih rendah. Pendekatan ini memastikan bahwa setiap sampel dari kelas minoritas memiliki pengaruh yang lebih besar selama proses pelatihan model, sehingga model dapat belajar dari kelas-kelas yang kurang representatif dengan lebih efektif

3.2 Hasil Pemodelan XGBoost

Model Extreme Gradient Boosting (XGBoost) pada penelitian ini dilatih menggunakan data pelatihan yang telah melalui tahap pra-pemrosesan, termasuk seleksi fitur geologi–geospasial dan penanganan ketidakseimbangan kelas menggunakan *sample weight*

3.2.1 Performa Model Prediksi XGBoost

Berdasarkan hasil evaluasi pada data training yang disajikan pada Tabel 6 dibawah.

Tabel 6. Performa Model pada Data Training

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0,97	0,93	0,98	0,95

Terlihat pada Tabel 6, model Extreme Gradient Boosting (XGBoost) menunjukkan performa yang sangat baik dalam mempelajari pola data fosil trilobita. Model mencapai nilai akurasi sebesar 0,97, yang mengindikasikan bahwa mayoritas data pelatihan berhasil diklasifikasikan dengan benar ke dalam periode geologi yang sesuai.Nilai precision sebesar 0,93 menunjukkan bahwa prediksi yang dihasilkan oleh model memiliki tingkat ketepatan yang tinggi, dengan jumlah kesalahan prediksi positif yang relatif rendah. Sementara itu, recall sebesar 0,98 menandakan bahwa model mampu mengenali hampir seluruh sampel dari setiap kelas pada data pelatihan, termasuk kelas dengan jumlah data yang lebih sedikit. Selanjutnya, nilai F1-score sebesar 0,95 menggambarkan keseimbangan yang baik antara precision dan recall, sehingga memperkuat indikasi bahwa model telah mempelajari karakteristik data pelatihan secara efektif. Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa model XGBoost memiliki kapasitas yang sangat baik dalam memodelkan hubungan antara fitur geologi–geospasial dan periode fosil trilobita pada tahap pelatihan, sehingga layak untuk dilanjutkan ke tahap evaluasi pada data pengujian

3.2.2 Evaluasi Model Prediksi dengan Data Test

Evaluasi model dilakukan pada data testing yang mempunyai 7823 sampel yang tidak melalui preprocesssing berupa penanganan data imbalance yang dijelaskan pada Tabel 7.

Tabel 7. Performa Model pada Data Test

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0,94	0,87	0,93	0,89

Setelah menunjukkan performa yang sangat baik pada data training, model Extreme Gradient Boosting (XGBoost) selanjutnya dievaluasi menggunakan data testing untuk menilai kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Hasil evaluasi pada data pengujian yang disajikan pada Tabel 7 menunjukkan bahwa model mempertahankan kinerja yang baik dengan nilai akurasi sebesar 0,94, yang mengindikasikan bahwa sebagian besar sampel uji dapat diklasifikasikan dengan tepat ke dalam periode geologi fosil trilobita yang sesuai. Nilai precision sebesar 0,87 menunjukkan adanya sedikit peningkatan kesalahan prediksi positif dibandingkan pada data pelatihan, namun masih berada dalam rentang yang dapat diterima. Hal ini mencerminkan tantangan model dalam membedakan beberapa periode geologi yang memiliki karakteristik fitur yang saling tumpang tindih. Sementara itu, recall sebesar 0,93 mengindikasikan bahwa model tetap mampu mengenali mayoritas sampel fosil dari setiap kelas periode pada data pengujian. Nilai F1-score sebesar 0,89 menunjukkan keseimbangan yang cukup baik antara precision dan recall pada data pengujian. Perbedaan performa antara data pelatihan dan data pengujian bersifat wajar dan mengindikasikan bahwa model tidak mengalami overfitting yang signifikan. Dengan demikian, hasil ini menunjukkan bahwa model XGBoost memiliki kemampuan generalisasi yang baik dan stabil dalam memprediksi periode geologi fosil trilobita berdasarkan fitur geologi–geospasial yang digunakan.

3.2.3 Performa Per Class

Berdasarkan hasil dari performa Per Class yang tertera pada Tabel 8.

Tabel 8. Performa Per-Class pada Data Testing

	Precision	Recall	F1-Score	Support
Cambrian	0,95	0,97	0,95	2447

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

	Precision	Recall	F1-Score	Support
Carboniferous	0,75	0,94	0,83	254
Devonian	0,94	0,92	0,93	1109
Ordovician	0,98	0,92	0,95	3429
Permian	0,81	0,86	0,83	128
Silurian	0,80	0,95	0,87	456

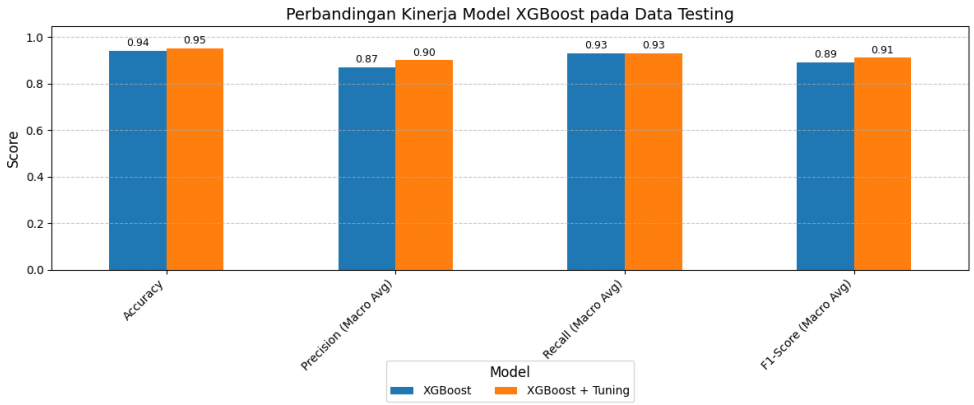
Analisis performa per-class pada data testing menunjukkan variasi kinerja model XGBoost dalam memprediksi periode geologi fosil trilobita. Periode Cambrian dan Ordovician mencapai performa tertinggi dengan F1-score 0,95, didukung oleh jumlah data yang besar, sehingga pola fitur geologi–geospasial dapat dipelajari secara lebih stabil dan representatif. Periode Devonian juga menunjukkan performa tinggi dengan F1-score 0,93, mengindikasikan bahwa karakteristik geologi pada periode ini dapat dikenali dengan baik oleh model.Sebaliknya, periode Carboniferous dan Permian memiliki performa yang lebih rendah (F1-score 0,83), yang kemungkinan disebabkan oleh jumlah sampel yang relatif sedikit serta kemiripan karakteristik geologi antar periode Paleozoikum akhir. Periode Silurian menunjukkan performa menengah dengan F1-score 0,87, ditandai oleh recall yang tinggi namun precision yang lebih rendah, yang mengindikasikan adanya tumpang tindih karakteristik dengan periode tetangga. Secara keseluruhan, hasil ini menegaskan bahwa performa model sangat dipengaruhi oleh jumlah data dan tingkat keunikan fitur geologi–geospasial pada masing-masing periode.

3.3 Hasil Tuning Hyperparameter

Peningkatan akurasi dan kemampuan generalisasi model XGBoost menjadi aspek penting dalam penelitian ini, sehingga dilakukan proses hyperparameter tuning menggunakan metode randomized search CV untuk memperoleh konfigurasi parameter yang optimal yang tertera pada Tabel 9 dan Gambar 3.

Tabel 9. Perbandingan Metrik Model

Metrik	XGBoost	XGBoost + Tuning
Accuracy	0.94	0.95
Precision	0.87	0.90
Recall	0.93	0.93
F1-Score	0.89	0.91



Gambar 3. Perbandingan Kinerja XGBoost

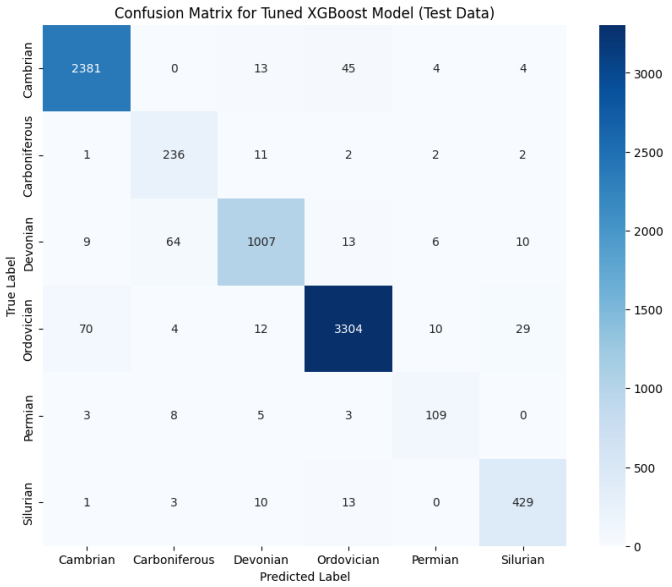
Berdasarkan Tabel 9 dan Gambar 3 perbandingan kinerja model, dapat dilihat bahwa penerapan hyperparameter tuning pada algoritma XGBoost memberikan peningkatan performa yang konsisten pada data pengujian. Model XGBoost tanpa tuning menghasilkan akurasi sebesar 0,94, sedangkan setelah dilakukan tuning akurasinya meningkat menjadi 0,95. Peningkatan ini menunjukkan bahwa konfigurasi parameter yang lebih optimal mampu memperbaiki kemampuan model dalam mengklasifikasikan periode fosil trilobita secara keseluruhan. Peningkatan paling signifikan terlihat pada metrik precision, yang naik dari 0,87 menjadi 0,90. Hal ini mengindikasikan bahwa proses tuning berhasil mengurangi jumlah false positive, sehingga prediksi model menjadi lebih tepat dan tidak terlalu agresif dalam mengklasifikasikan fosil ke periode tertentu. Sementara itu, nilai recall tetap berada pada angka 0,93 baik sebelum maupun sesudah tuning, yang menunjukkan bahwa kemampuan model dalam mengenali data dari masing-masing periode fosil tetap terjaga dengan baik dan tidak mengalami penurunan sensitivitas. Selain itu, nilai F1-score meningkat dari 0,89 menjadi 0,91 setelah tuning dilakukan. Kenaikan F1-score ini mencerminkan keseimbangan yang lebih baik antara precision dan recall, terutama dalam konteks klasifikasi multikelas dengan distribusi data yang tidak sepenuhnya seimbang. Dengan demikian, hyperparameter tuning berperan penting dalam memperbaiki stabilitas dan kualitas prediksi model tanpa menimbulkan overfitting yang signifikan. Secara keseluruhan, hasil ini menegaskan bahwa penerapan hyperparameter tuning menggunakan RandomizedSearchCV mampu mengoptimalkan performa model XGBoost, khususnya dalam meningkatkan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

ketepatan prediksi, sehingga model menjadi lebih andal untuk digunakan dalam prediksi periode fosil trilobita berbasis data geologi dan geospasial.

3.2.1 Evaluasi Confusion Matrix

Berdasarkan hasil dari Confusion Matrix yang tervisualisasi pada Gambar 4.



Gambar 4. Confusion Matrix

Confusion matrix yang ditunjukan pada Gambar 4 menunjukan bahwa model XGBoost yang telah dituning mampu melakukan klasifikasi periode fosil trilobita dengan baik, yang ditandai oleh dominasi nilai pada diagonal utama sebagai indikasi prediksi yang benar. Kelas Cambrian dan Ordovician memiliki performa tertinggi dengan jumlah true positive yang sangat besar, menunjukkan pola fitur geologi–geospasial yang jelas dan mudah dikenali model. Sebagian kesalahan prediksi umumnya terjadi pada periode geologi yang berdekatan, seperti Devonian, Carboniferous, dan Silurian, yang kemungkinan disebabkan oleh kemiripan karakteristik lingkungan dan litologi antar periode. Kelas dengan jumlah data lebih sedikit, seperti Permian, masih menunjukkan performa yang cukup baik meskipun lebih rentan terhadap misklasifikasi. Secara keseluruhan, hasil ini menegaskan bahwa hyperparameter tuning meningkatkan kemampuan generalisasi model, dengan kesalahan prediksi yang masih bersifat logis secara geologis

3.4 Analisis Limitasi dan Implikasi Penelitian

Hasil penelitian menunjukkan bahwa model XGBoost memberikan performa prediksi yang sangat baik pada periode Cambrian dan Ordovician. Secara geologis, hal ini dapat dijelaskan oleh karakteristik litologi dan lingkungan pengendapan pada kedua periode tersebut yang relatif khas, sehingga pola geologi yang terbentuk lebih mudah dikenali oleh model. Periode Cambrian didominasi oleh sedimen laut dangkal awal, sedangkan Ordovician menunjukkan diversifikasi lingkungan laut dengan ciri litologi yang lebih bervariasi. Misklasifikasi terutama terjadi pada periode yang berdekatan secara stratigrafi, seperti antara Silurian dan Ordovician. Kondisi ini wajar karena transisi biostratigrafi antar kedua periode tersebut bersifat gradual dan memiliki kemiripan karakteristik geologi. Dengan demikian, kesalahan klasifikasi yang terjadi mencerminkan kedekatan geologis antar periode, bukan kelemahan model secara keseluruhan.

Meskipun pendekatan yang diusulkan dalam penelitian ini menunjukkan performa yang baik dalam memprediksi periode geologi fosil trilobita, terdapat beberapa limitasi yang perlu dipertimbangkan dalam konteks pengembangan dan penerapan model lebih lanjut.

Limitasi pertama terkait dengan ketergantungan model pada kualitas dan kelengkapan data. Dataset yang digunakan berasal dari sumber sekunder (Kaggle) dan mengandalkan kompilasi berbagai catatan paleontologi dengan tingkat ketelitian geospasial dan geologi yang bervariasi. Ketidakpastian dalam penentuan lokasi, litologi, maupun konteks lingkungan pengendapan pada sebagian entri berpotensi memengaruhi akurasi prediksi periode fosil.

Limitasi kedua berkaitan dengan ruang lingkup fitur yang digunakan. Penelitian ini secara sengaja membatasi fitur pada atribut geologi–geospasial utama, yaitu longitude, latitude, lithology, environment, dan collection_type. Meskipun pendekatan ini efektif dalam mengurangi noise dan meningkatkan generalisasi model, pembatasan tersebut juga berarti bahwa faktor taksonomi yang lebih rinci, seperti variasi genus atau morfologi spesifik trilobita, belum sepenuhnya dimanfaatkan. Pada kasus tertentu, informasi morfologis dapat mengandung sinyal temporal yang kuat, sehingga penggabungan fitur morfologi berpotensi meningkatkan performa prediksi.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

Limitasi ketiga berkaitan dengan ketidakseimbangan distribusi kelas periode geologi. Meskipun penanganan imbalance telah dilakukan melalui pemberian sample weight, pendekatan ini tetap bergantung pada asumsi bahwa bobot kelas mampu merepresentasikan kompleksitas distribusi data yang sebenarnya. Pada periode dengan jumlah data yang sangat sedikit, seperti Permian atau Carboniferous, model masih menunjukkan performa yang lebih rendah dibanding periode dengan jumlah sampel besar. Hal ini mengindikasikan bahwa keterbatasan data historis tetap menjadi tantangan utama yang tidak sepenuhnya dapat diatasi melalui teknik pembobotan.

Limitasi keempat berkaitan dengan cakupan validasi model. Evaluasi pada penelitian ini berfokus pada metrik teknis klasifikasi seperti akurasi, precision, recall, dan F1-score. Walaupun metrik tersebut cukup untuk menilai performa komputasional, penelitian ini belum melibatkan validasi ahli (expert validation) dari paleontolog untuk menilai kesesuaian prediksi model dengan interpretasi stratigrafi dan biostratigrafi yang digunakan dalam praktik ilmiah. Keterlibatan ahli domain berpotensi memberikan perspektif tambahan mengenai nilai ilmiah dan interpretabilitas hasil prediksi. Dari sisi implikasi praktis, metode yang diusulkan memiliki potensi aplikasi yang menjanjikan sebagai alat bantu analisis paleontologi digital, khususnya dalam eksplorasi awal penentuan periode fosil pada dataset berskala besar. Kemampuan model dalam mengenali pola geologi–geospasial yang konsisten menunjukkan bahwa pendekatan ini dapat digunakan sebagai sistem pendukung keputusan (decision support system) untuk membantu peneliti memfilter atau mengelompokkan fosil berdasarkan estimasi periode geologinya. Namun demikian, model ini belum dimaksudkan untuk menggantikan penentuan stratigrafi manual secara penuh, melainkan sebagai pelengkap yang dapat meningkatkan efisiensi dan objektivitas analisis.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi periode umur fosil trilobita menggunakan algoritma Extreme Gradient Boosting (XGBoost) dengan seleksi fitur geologi–geospasial dan optimasi hyperparameter. Permasalahan utama yang dihadapi meliputi kompleksitas data paleontologi, keberadaan nilai hilang, serta ketidakseimbangan distribusi kelas pada variabel target time_period yang berpotensi menurunkan akurasi prediksi. Melalui tahapan preprocessing yang sistematis, meliputi pembersihan data, imputasi nilai hilang, encoding fitur kategorikal, pembagian data menggunakan stratified train–test split, serta penanganan class imbalance menggunakan class weight adjustment, diperoleh dataset yang lebih representatif untuk pemodelan. Seleksi fitur longitude, latitude, lithology, environment, dan collection_type terbukti efektif dalam merepresentasikan faktor pembentukan umur fosil.Model XGBoost tanpa hyperparameter tuning menghasilkan performa awal dengan akurasi sebesar 88% dan F1-score sebesar 84%, yang menunjukkan kemampuan dasar model namun masih memiliki keterbatasan dalam generalisasi. Setelah dilakukan optimasi hyperparameter menggunakan RandomizedSearchCV, performa model mengalami peningkatan signifikan dengan akurasi sebesar 95%, precision sebesar 90%, recall sebesar 93%, dan F1-score sebesar 91% pada data pengujian. Peningkatan ini menunjukkan bahwa hyperparameter tuning berperan penting dalam mengoptimalkan kompleksitas model dan meningkatkan kemampuan prediksi pada data yang belum pernah dilihat sebelumnya. Meskipun demikian, penelitian ini memiliki keterbatasan pada penggunaan satu sumber dataset dan jumlah fitur yang relatif terbatas, sehingga penelitian selanjutnya disarankan untuk mengintegrasikan fitur tambahan, menggunakan dataset yang lebih luas, serta mengeksplorasi pendekatan model dan teknik penanganan imbalance lainnya guna meningkatkan akurasi dan generalisasi model.

REFERENCES

[1] J. D. Holmes and G. E. Budd, “Reassessing a cryptic history of early trilobite evolution,” *Commun Biol*, vol. 5, no. 1, Dec. 2022, doi: 10.1038/s42003-022-04146-6.

[2] C. S. Marques, E. Malafaia, S. Pereira, V. F. Santos, and E. Dufourq, “A review of machine learning applications for identification and classification problems in paleontology,” , *Elsevier B.V.*, vol. 91, Nov. 01, 2025doi: 10.1016/j.ecoinf.2025.103329.

[3] B. T. Kopperud, S. Lidgard, and L. H. Liow, “Text-mined fossil biodiversity dynamics using machine learning,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 286, no. 1901, Apr. 2019, doi: 10.1098/rspb.2019.0022.

[4] J. Castle-Jones *et al.*, “Integrated biostratigraphy, chemostratigraphy and geochronology of the lower Cambrian succession in the western Stansbury Basin, South Australia,” *Australian Journal of Earth Sciences*, vol. 72, no. 2, pp. 182–212, 2025, doi: 10.1080/08120099.2025.2473098.

[5] E. Hodgson, J. McCoy, K. Webber, N. Nuñez Otaño, J. O’Keefe, and M. Pound, “A global dataset of fossil fungi records from the Cenozoic,” *Scientific Data* , vol. 12, no. 1, Dec. 2025, doi: 10.1038/s41597-025-04553-4.

[6] D. A. Dinanthi, E. Ramadanti, C. Sri, K. Aditya, and D. R. Chandranegara, “Diabetes Detection Using Extreme Gradient Boosting (XGBoost) with Hyperparameter Tuning,” *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 78–84, May 2024, doi: 10.35882/ijeeemi.v6i2.351.

[7] M. Kang, K. Pham, K. Kwon, S. Yang, and H. Choi, “A Hybrid Numerical-ML Model for Predicting Geological Risks in Tunneling with Electrical Methods,” *KSCE Journal of Civil Engineering*, vol. 28, no. 12, pp. 5972–5986, Dec. 2024, doi: 10.1007/s12205-024-0066-z.

[8] J. Bahn, G. H. Alférez, and K. Snyder, “Machine Learning Classification of Fossilized Pectinodon bakkeri Teeth Images: Insights into Troodontid Theropod Dinosaur Morphology,” *Mach Learn Knowl Extr*, vol. 7, no. 2, Jun. 2025, doi: 10.3390/make7020045.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2181−2192
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8862

[9] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, "Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures," *Front Earth Sci (Lausanne)*, vol. 9, Oct. 2021, doi: 10.3389/feart.2021.761990.
[10] J. Zhang, R. Wang, A. Jia, and N. Feng, "Optimization and Application of XGBoost Logging Prediction Model for Porosity and Permeability Based on K-means Method," *Applied Sciences (Switzerland)*, vol. 14, no. 10, May 2024, doi: 10.3390/app14103956.
[11] N. M. Shahani, M. Kamran, X. Zheng, C. Liu, and X. Guo, "Application of gradient boosting machine learning algorithms to predict uniaxial compressive strength of soft sedimentary rocks at Thar coalfield," *Advances in Civil Engineering*, vol. 2021, Oct. 2021, doi: 10.1155/2021/2565488.
[12] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, "Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures," *Front Earth Sci (Lausanne)*, vol. 9, Oct. 2021, doi: 10.3389/feart.2021.761990.
[13] A. A. Syahputra and R. E. Saputro, "Application of the XGBoost Model with Hyperparameter Tuning for Industry Classification for Job Applicants," *sinkron*, vol. 8, no. 3, pp. 1920–1931, Jul. 2024, doi: 10.33395/sinkron.v8i3.13840.
[14] A. Dendi Rachmatsyah, T. Sugihartono, and K. Irfan, "Perbandingan Teknik Optimasi Grid Search dan Randomized Search dalam Meningkatkan Akurasi Metode Klasifikasi SVM Pada Sentimen Ulasan Pengguna Aplikasi JKN Mobile," *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 8, no. 1, pp. 13–22, Jan. 2025, doi: https://doi.org/10.36080/skanika.v8i1.3328.
[15] A. Jarmakovica, "Machine learning-based strategies for improving healthcare data quality: an evaluation of accuracy, completeness, and reusability," *Front Artif Intell*, vol. 8, Jul. 2025, doi: 10.3389/frai.2025.1621514.
[16] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: A system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63–72, Mar. 2022, doi: 10.1016/j.future.2021.09.033.
[17] W. Wang, C. Xue, J. Zhao, C. Yuan, and J. Tang, "Machine learning-based field geological mapping: A new exploration of geological survey data acquisition strategy," *Ore Geol Rev*, vol. 166, Mar. 2024, doi: 10.1016/j.oregeorev.2024.105959.
[18] D. Breskuvien and G. Dzemyda, "Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions," *International Journal of Computers, Communications and Control*, vol. 18, no. 3, Jun. 2023, doi: 10.15837/ijccc.2023.3.5433.
[19] W. Albattah and R. U. Khan, "Impact of imbalanced features on large datasets," *Front Big Data*, vol. 8, Mar. 2025, doi: 10.3389/fdata.2025.1455442.
[20] Z. Wang, X. Chu, D. Li, H. Yang, and W. Qu, "Cost-sensitive matrixized classification learning with information entropy," *Appl Soft Comput*, vol. 116, Feb. 2022, doi: 10.1016/j.asoc.2021.108266.
[21] J. Han, K. Shu, and Z. Wang, "Predicting energy use in construction using Extreme Gradient Boosting," *PeerJ Comput Sci*, vol. 9, Aug. 2023, doi: 10.7717/peerj-cs.1500.
[22] F. Nurrahman, H. Wijayanto, A. H. Wigena, and N. Nurjanah, "PRE-PROCESSING DATA ON MULTICLASS CLASSIFICATION OF ANEMIA AND IRON DEFICIENCY WITH THE XGBOOST METHOD," *Barekeng*, vol. 17, no. 2, pp. 767–774, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0767-0774.
[23] Sugiarto *et al.*, "Optimizing The XGBoost Model with Grid Search Hyperparameter Tuning for Maximum Temperature Forecasting," *Journal of Applied Data Sciences*, vol. 6, no. 4, pp. 2517–2529, Dec. 2025, doi: 10.47738/jads.v6i4.885.
[24] H. Wijaya, D. P. Hostiadi, and E. Triandini, "Optimization XGBoost Algorithm Using Parameter Tunning in Retail Sales Prediction," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 13, no. 3, Dec. 2024, doi: 10.23887/janapati.v13i3.82214.
[25] C. G. L. Pringandana and K. Kusnawi, "A Comparative Analysis of Hyperparameter-Tuned XGBoost and LightGBM for Multiclass Rainfall Classification in Jakarta," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 2467–2483, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4965.
[26] D. S. Soper, "Greed is good: Rapid hyperparameter optimization and model selection using greedy k-fold cross validation," *Electronics (Switzerland)*, vol. 10, no. 16, Aug. 2021, doi: 10.3390/electronics10161973.
[27] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," *Ecol Monogr*, vol. 93, no. 1, Feb. 2023, doi: 10.1002/ecm.1557.
[28] D. Liang, X. Jin, Y. Yuan, and R. Zou, "Performance Analysis of Machine Learning Methods," in *Journal of Physics: Conference Series*, Institute of Physics, vol. 2023, Oct. 2023. doi: 10.1088/1742-6596/2428/1/012039.
[29] I. Imantoko, A. Hermawan, and D. Avianto, "Comparative analysis of support vector machine and k-nearest neighbors with a pyramidal histogram of the gradient for sign language detection," *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 11, no. 2, pp. 107–118, Jul. 2021, doi: 10.31940/matrix.v11i2.2433.
[30] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.
[31] R. Irmanita, S. S. Prasetiyowati, and Y. Sibaroni, "Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naïve Bayes," *Jurnal RESTI*, vol. 5, no. 1, pp. 10–16, Feb. 2021, doi: 10.29207/resti.v5i1.2770.