BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

Integrasi Augmentasi Data dan Machine Learning dalam Prediksi Magnitudo Gempa Bumi

Hastari Utama[1], Ahlihi Masruro[2], Toto Indriyatmoko[3], Sudarmanto[4]
[1,2,3]Program Studi Teknik Informatik, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
[4]Program Studi Rekayasa Perangkat Lunak Aplikasi, Fakultas Teknologi Informasi, Universitas Teknologi Digital Indonesia
[1,2,3]Jl. Ring Road Utara, Ngringin, Condong Catur, Depok, Sleman, telp. (0274) 884201 – 207
[1,2,3]Jl. Raya Janti Karang Jambe No. 143, Banguntapan, Bantul, telp. (0274) 486664

Corresponding e-mail: [1]utama@amikom.ac.id

Abstrak

Penelitian ini bertujuan meningkatkan akurasi prediksi magnitudo gempa bumi melalui integrasi teknik augmentasi data dan algoritma Random Forest Regressor, dengan dukungan visualisasi geospasial untuk analisis mendalam. Dataset diperoleh dari USGS (United States Geological Survey) dalam format CSV, mencakup lebih dari seribu peristiwa gempa bumi global selama satu bulan terakhir, dengan parameter seperti lokasi (latitude, longitude), kedalaman, magnitudo, dan kualitas pencatatan. Mengingat distribusi data yang tidak seimbang, dengan dominasi gempa kecil dan jarangnya gempa besar, dilakukan augmentasi berbasis noise injection pada fitur spasial dan kedalaman. Proses ini menghasilkan dataset baru yang lima kali lebih besar dari data asli tanpa mengubah integritas geografis. Evaluasi kinerja menunjukkan peningkatan signifikan: MAE turun 57,6% (0,2467 menjadi 0,1046), RMSE berkurang 46,6% (0,3499 menjadi 0,1868), MSE menurun 71,5% (0,1225 menjadi 0,0349), dan R^2 meningkat dari 0,9493 menjadi 0,9817. Visualisasi geospasial mengungkap pola sebaran gempa dan hotspot di wilayah aktif tektonik seperti Pasifik Ring of Fire, California, Alaska, dan Indonesia. Hasil ini membuktikan bahwa augmentasi data merupakan strategi krusial untuk meningkatkan generalisasi model dan prediksi, dengan potensi integrasi pada sistem peringatan dini serta pemetaan risiko seismik.

Kata kunci: *Augmentasi Data, Gempa Bumi, Prediksi Magnetudo, Random Forest Regressor, USGS, Visualisasi Data.*

Abstract

This study aims to enhance earthquake magnitude prediction accuracy by integrating data augmentation techniques with a Random Forest Regressor, supported by geospatial visualization for comprehensive analysis. The dataset, sourced from the official USGS (United States Geological Survey) in CSV format, contains over a thousand global earthquake events recorded in the past month, including spatial (latitude, longitude), depth, magnitude, and quality parameters. Given the highly imbalanced distribution dominated by small earthquakes with few large events, noise injection-based augmentation was applied to spatial features and depth, expanding the training data to five times its original size while maintaining geographic and physical integrity. Model evaluation demonstrated significant improvements: MAE decreased from 0.2467 to 0.1046 (57.6%), RMSE from 0.3499 to 0.1868 (46.6%), MSE from 0.1225 to 0.0349 (71.5%), and R^2 increased from 0.9493 to 0.9817. Geospatial visualization revealed temporal and spatial patterns concentrated in tectonically active regions such as the Pacific Ring of Fire, California, Alaska, and Indonesia. These findings confirm that data augmentation is not merely supplementary but a vital strategy for improving model generalization and predictive performance, particularly for underrepresented large-magnitude earthquakes. This research provides both scientific and practical contributions for disaster mitigation, with strong potential for integration into early warning systems and seismic risk mapping.

Keywords: *Data Augmentation, Earthquake, Magnitude Prediction, Random Forest Regressor, USGS, Data Visualization.*

1. Pendahuluan

Gempa bumi merupakan salah satu fenomena alam paling dahsyat dan berpotensi merusak yang dapat terjadi di berbagai belahan dunia, terutama di wilayah yang berada di sekitar lempeng tektonik aktif seperti Cincin Api Pasifik(Aden-Antoniów et al., 2022; Zhang et al., 2023). Peristiwa gempa bumi tidak hanya menimbulkan kerusakan fisik pada infrastruktur, tetapi juga berdampak besar terhadap kehidupan manusia, ekonomi, dan lingkungan secara menyeluruh(Saleem & Rashid, 2023). Dengan semakin meningkatnya populasi di daerah rawan gempa, penting untuk mengembangkan sistem deteksi dini dan prediksi yang akurat guna meminimalkan risiko dan korban jiwa. Data historis tentang gempa bumi sangat krusial dalam memahami pola aktivitas seismik dan membangun model prediktif yang handal. Oleh karena itu, penelitian ini bertujuan untuk memanfaatkan data gempa bumi dari sumber resmi seperti *USGS* (*United States Geological Survey*) sebagai dasar analisis(Ghimire et al., 2022).

Salah satu tantangan utama dalam prediksi gempa bumi adalah ketidakseimbangan dan keterbatasan jumlah data, terutama untuk peristiwa dengan magnitudo tinggi yang relatif jarang terjadi(Hakak et al., 2021; Liu et al., 2023). Hal ini menyebabkan bias dalam pelatihan model machine learning, di mana model cenderung lebih baik dalam memprediksi gempa kecil daripada gempa besar yang memiliki dampak signifikan(D'Amico et al., 2022). Untuk mengatasi masalah ini, teknik augmentasi data menjadi strategi penting yang dapat digunakan untuk memperkaya dataset dengan menciptakan sampel sintetis yang realistis namun tetap sesuai dengan pola geologis yang ada. Augmentasi data membantu meningkatkan generalisasi model dan mencegah *overfitting*, sehingga model menjadi lebih robust terhadap variasi data nyata(Chelidze et al., 2022).

Dalam konteks machine learning, *Random Forest Regressor* dipilih sebagai metode prediksi karena kemampuannya dalam menangani data non-linear, mengurangi *overfitting* melalui ensemble learning, serta memberikan informasi penting mengenai fitur-fitur yang paling berpengaruh terhadap hasil prediksi(Mori et al., 2022; Zheng et al., 2022). Model ini cocok untuk tugas regresi seperti memprediksi magnitudo gempa berdasarkan parameter geospasial dan seismik seperti kedalaman, lokasi (lintang dan bujur), jumlah stasiun pencatat, dan jarak ke pusat gempa(Sharma et al., 2022). Selain itu, Random Forest juga dapat bekerja dengan baik bahkan ketika terdapat noise atau outlier dalam data. Kondisi umum pada data gempa bumi(Jamjoom & El Hindi, 2016). Implementasi model ini akan dilakukan dengan pengujian performa menggunakan metrik seperti RMSE, MAE, dan R^2 untuk menilai akurasi prediksi(Bilal et al., 2022). Visualisasi data merupakan komponen penting dalam proses eksplorasi dan interpretasi temuan, terutama dalam studi geospasial seperti ini(Witten et al., 2011). Dengan menggunakan peta interaktif berbasis GIS (seperti Folium atau Plotly), peneliti dapat menampilkan distribusi gempa bumi secara spasial dan temporal, mengidentifikasi hotspot seismik, serta memahami hubungan antara lokasi dan intensitas gempa(Banna et al., 2020). Visualisasi juga membantu dalam menyampaikan hasil kepada pemangku kepentingan, termasuk instansi mitigasi bencana, pemerintah daerah, dan masyarakat umum. Melalui representasi grafis yang informatif, penelitian ini tidak hanya menghasilkan model prediksi, tetapi juga menyediakan alat komunikasi yang efektif untuk pengambilan keputusan berbasis data.

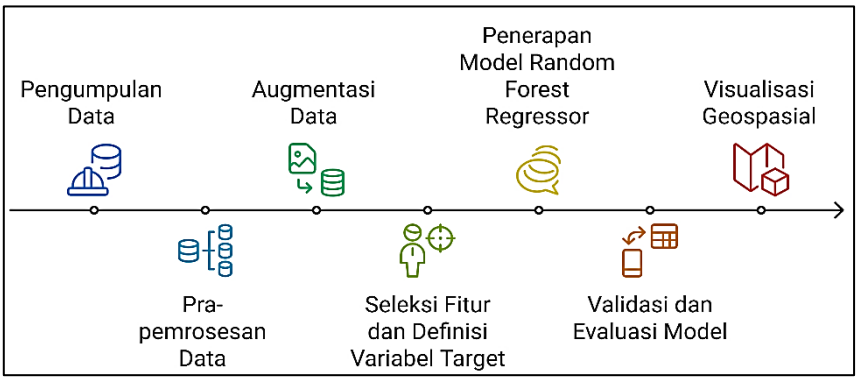
Berdasarkan latar belakang tersebut, penelitian ini dirancang untuk mengintegrasikan teknik augmentasi data, pendekatan machine learning dengan Random Forest Regressor, dan visualisasi geospasial dalam satu kerangka kerja komprehensif guna meningkatkan akurasi dan relevansi prediksi gempa bumi. Tujuan utamanya adalah menghasilkan model yang tidak hanya akurat secara statistik, tetapi juga berguna secara praktis dalam konteks mitigasi bencana. Penelitian ini diharapkan dapat memberikan kontribusi ilmiah dan aplikatif bagi pengembangan sistem peringatan dini gempa bumi, serta menjadi referensi bagi penelitian lanjutan di bidang seismologi dan pembelajaran mesin. Dengan pendekatan multidisiplin ini, diharapkan dapat memperkuat ketahanan sosial dan infrastruktur di wilayah rawan gempa.

2. Metode Penelitian

Penelitian ini dirancang untuk mengintegrasikan pendekatan data-driven dalam memprediksi magnitudo gempa bumi secara akurat dan dapat diandalkan, dengan memanfaatkan potensi teknologi machine learning serta strategi penguatan data untuk mengatasi tantangan utama dalam analisis seismik(Ruz et al., 2020; Sugiyama et al., 2021). Data penelitian diperoleh dari USGS (United States Geological Survey) melalui feed CSV yang memuat informasi gempa bumi global selama satu bulan terakhir, mencakup lokasi (lintang, bujur), kedalaman, magnitudo, dan kualitas pencatatan. Tahap awal meliputi pengumpulan dan pra-pemrosesan data, termasuk

BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

penyaringan jenis peristiwa "earthquake" serta penanganan nilai kosong. Seleksi fitur difokuskan pada variabel yang berpengaruh signifikan terhadap magnitudo. Mengingat distribusi data yang tidak seimbang, dengan gempa besar jauh lebih sedikit, diterapkan teknik augmentasi untuk memperkuat representasi kelas minoritas dan mengurangi bias, sehingga model prediksi menjadi lebih robust dan mampu melakukan generalisasi lebih baik. Dalam konteks algoritma machine learning, Random Forest Regressor dipilih karena kemampuannya menangani data non-linear, mengurangi overfitting melalui ensemble learning, serta memberikan interpretasi terhadap pentingnya fitur-fitur dalam prediksi(Chelidze et al., 2022; Manral & Chaudhary, 2023) dan hasilnya lebih efektif (Sharma et al., 2022) yang disajikan pada Gambar 1.



Gambar 1. Alur Penelitian Prediksi Magnitudo Gempa

Metode Penelitian memberikan penjelasan tentang langkah-langkah, data, lokasi penelitian, metode evaluasi yang digunakan. Langkah-langkah ini ditunjukkan pada Gambar 1 dimana penjelasan terstruktur tentang algoritma atau metode dari penelitian dijelaskan pada sub bab berikutnya.

2.1. Pengumpulan dan Pra-pemrosesan Data

Data gempa bumi dikumpulkan dari sumber resmi *USGS* melalui URL: https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php . Dataset berupa file CSV bernama all_month.csv yang mencakup peristiwa gempa bumi selama satu bulan terakhir, dengan 24 kolom informasi. Langkah pertama dalam proses analisis adalah melakukan *data cleaning* atau pembersihan data, yaitu menghapus baris yang tidak relevan seperti *explosion*, *quarry blast*, atau *landslide* dengan memfilter hanya pada baris yang memiliki nilai type earthquake. Selain itu, dilakukan penanganan terhadap nilai-nilai kosong (*missing values*) pada kolom-kolom penting seperti mag, latitude, longitude, dan depth. Kolom-kolom yang tidak digunakan dalam model prediksi diabaikan untuk menghindari noise seperti pada Tabel 1.

Tabel 1. Karakteristik Dataset

Karakteristik	Deskripsi
Sumber Data	United States Geological Survey (USGS) – Earthquake Hazards Program
URL Sumber	https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php
Format File	CSV (Comma-Separated Values), ASCII text
Jenis Data	Data gempa bumi global, mencakup peristiwa seismik aktif selama satu bulan terakhir
Periode Waktu	Berdasarkan data dalam file:14 Juli 2025 – 31 Juli 2025(dari waktutime)
Jumlah Baris (Data Observasi)	~1.000–1.500 baris (berdasarkan jumlah entri dalam file)
Jumlah Kolom (Fitur)	24 kolom sesuai spesifikasi USGS
Kolom Utama	time,latitude,longitude,depth,mag,magType,nst,gap,dmin,rms,net,id, updated,place,type,locationSource,magSource,horizontalError,depth Error,magError,magNst,status
Target Prediksi	Magnitudo gempa (mag) – variabel numerik kontinu

Karakteristik	Deskripsi
Tipe Variabel	Campuran: Numerik (float/int), Kategorikal (string), Tanggal (datetime)
Rentang Nilai mag	Dari0.17hingga14.09(gempa besar di Timor Leste & Indonesia)
Rata-rata Magnitudo	Sekitar2.6–3.0(kebanyakan gempa kecil)
Distribusi Magnitudo	Cenderung tidak seimbang: banyak gempa kecil (magnitudo < 3), sedikit gempa besar (≥ 5).
Wilayah Geografis Tersebar	Global, dengan fokus tinggi di wilayah aktif tektonik: a) California, Nevada, Washington (AS) b) Alaska, Rusia Timur (Petropavlovsk-Kamchatsky) c) Jepang d) Indonesia (Banda Sea, Sulawesi) e) Hawaii f) Chile, Timor Leste, Tanzania g) Solok Islands, Mauritius
Kualitas Data	Tinggi untuk gempa besar; beberapa nilai kosong padamag,magError,depthError, dll.
Status Data	Campuran:automatic(otomatis),reviewed(telah direview oleh ahli)
Sumber Jaringan Pencatat (net)	us(USGS),nc(Caltech),nn(Nevada),ci(California),uw(Washington),hv(Hawaii),ak(Alaska),uu(Utah),tx(Texas),ok(Oklahoma),av(Alaska), mb(Montana), dll.
Metode Pengukuran Magnitudo	Beragam:ml(magnitude lokal),md(magnitude depth),mb(body wave),mww(moment magnitude),mb(mantle wave)
Ketidakpastian Data	Beberapa kolom sepertimagError,depthError,horizontalErrormenyediakan informasi ketidakpastian pengukuran.
Catatan Spesial	Terdapat entri non-gempa seperti*explosion*(ledakan tambang),*quarry blast*, yang harus di-filter jika hanya ingin analisis gempa alami.

Tabel 1 memuat data peristiwa seismik global selama satu bulan dengan 24 kolom, mencakup waktu, koordinat, kedalaman, magnitudo, jenis skala, dan parameter kualitas pencatatan. Distribusi tidak seimbang, didominasi gempa kecil (<3) dengan sedikit gempa besar (≥5), terkonsentrasi di wilayah aktif tektonik. Dataset juga memuat entri non-seismik yang perlu difilter, dan cocok untuk analisis ilmiah, visualisasi spasial, serta penerapan machine learning untuk prediksi magnitudo.

2.2. Seleksi Fitur dan Definisi Variabel Target

Variabel target dalam penelitian ini adalah magnitudo gempa bumi (*mag*), karena merupakan parameter utama untuk menilai kekuatan gempa dan menjadi tujuan utama prediksi. Fitur-fitur yang digunakan sebagai input model meliputi: koordinat geografis (lintang dan bujur), kedalaman pusat gempa (*depth*), jumlah stasiun pencatat (*nst*), jarak minimum ke stasiun terdekat (*dmin*), sudut antar stasiun (gap), RMS residual waktu (rms), serta parameter lain seperti horizontalError, depthError, dan magError. Semua fitur numerik dimaksudkan untuk memberikan informasi spasial, temporal, dan seismik yang relevan bagi prediksi magnitudo.

2.3. Augmentasi Data

Dalam penelitian ini, teknik augmentasi data diterapkan untuk mengatasi masalah ketidakseimbangan (*imbalanced data*) yang umum terjadi pada dataset gempa bumi, di mana jumlah peristiwa dengan magnitudo tinggi (gempa besar) jauh lebih sedikit dibandingkan gempa kecil(Chelidze et al., 2022). Untuk memperkaya representasi data pelatihan dan meningkatkan generalisasi model prediksi, dilakukan augmentasi sintetis melalui pembuatan variasi data berbasis distribusi normal pada fitur spasial dan kedalaman(Varshney et al., 2023). Untuk mengatasi masalah ketidakseimbangan data dan meningkatkan robustness model prediksi, dilakukan augmentasi data sintetis berbasis noise injection pada fitur spasial dan kedalaman(Frith, 2023). Proses dimulai dari data asli (df), lalu dibuat empat salinan tambahan (n_aug = 4). Pada setiap salinan, nilai koordinat lintang (latitude) dan bujur (longitude) diberi gangguan acak dengan distribusi normal berstandar deviasi 0,01 derajat (setara ±1 km), sementara kedalaman pusat gempa

(depth) diberi noise dengan standar deviasi 0,5 km. Gangguan ini dirancang agar tetap sesuai dengan ketidakpastian alami dalam pengukuran seismik dan tidak menyalahi batasan geografis atau fisik. Seluruh dataset hasil augmentasi, termasuk data asli, kemudian digabungkan menjadi satu kesatuan menggunakan pandas.concat(), menghasilkan dataset baru dengan total jumlah sampel sebanyak lima kali lipat dari data asli. Hasil augmentasi ini digunakan sebagai input untuk pelatihan model Random Forest Regressor, dengan harapan bahwa penambahan variasi data akan membantu model belajar pola yang lebih umum dan meningkatkan akurasi prediksi, terutama untuk gempa besar yang langka.

2.4. Penerapan Model Random Forest Regressor

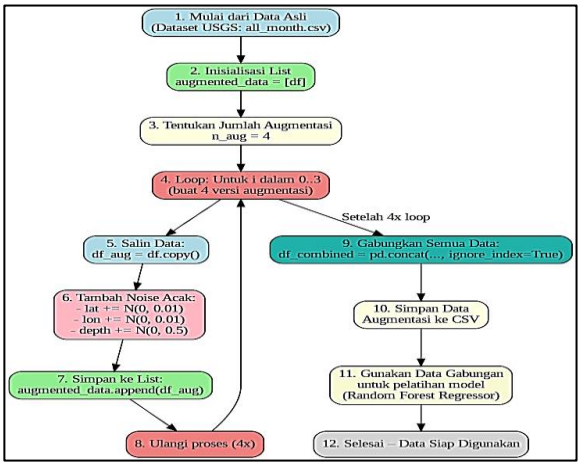
Model prediksi utama yang digunakan adalah Random Forest Regressor, sebuah algoritma ensemble learning yang robust terhadap overfitting dan mampu menangani data non-linear(Pasupulety et al., 2019). Model ini dibangun dengan menggunakan 100 hingga 300 pohon keputusan (*estimators*) dan mengoptimalkan parameter seperti max_depth, min_samples_split, dan n_estimators melalui cross-validation silang (k-fold CV, k=5). Model dilatih menggunakan data yang telah diproses dan diperkaya melalui augmentasi. Evaluasi performa model dilakukan menggunakan metrik-metrik regresi standar, yaitu Root Mean Square Error (RMSE), Mean Absolute Error (MAE), dan R^2 Score, untuk menilai akurasi prediksi magnitudo.

2.5. Validasi dan Evaluasi Model

Untuk menjamin keandalan model, dilakukan pembagian data menjadi dua bagian: training set (80%) dan test set (20%)(Zhao, 2017). Model dilatih pada data latih dan diuji pada data uji yang tidak pernah dilihat sebelumnya. Selain itu, dilakukan *cross-validation* untuk menghindari bias estimasi(Saputro & Sari, 2019). Hasil prediksi kemudian dibandingkan dengan nilai aktual magnitudo untuk menganalisis kesalahan sistematis, terutama pada gempa besar yang sulit diprediksi. Selain itu, dilakukan analisis *feature importance* dari Random Forest untuk mengetahui fitur mana yang paling berpengaruh terhadap prediksi magnitudo, seperti kedalaman, lokasi, atau jumlah stasiun(N & Thirukkumaran, 2020).

2.6. Visualisasi Geospasial

Sebagai bagian integral dari penelitian, visualisasi data dilakukan untuk memahami pola spasial dan temporal gempa bumi. Peta interaktif dibuat menggunakan library Python seperti Folium atau Plotly, menampilkan distribusi lokasi gempa bumi berdasarkan magnitudo dan waktu.



Gambar 2. Alur Augmentasi Data

Gempa besar ditandai dengan simbol lebih besar dan warna merah, sementara gempa kecil dengan warna hijau. Selain itu, digunakan heatmap untuk menunjukkan *hotspot* aktivitas seismik, plot time-series untuk tren frekuensi gempa harian, serta scatter plot antara kedalaman dan magnitudo untuk mengidentifikasi pola hubungan fisik(Fornasari et al., 2022). Visualisasi juga digunakan untuk menampilkan hasil prediksi model, termasuk kesalahan prediksi (residuals) dan distribusi error secara spasial disajikan pada Gambar 2.

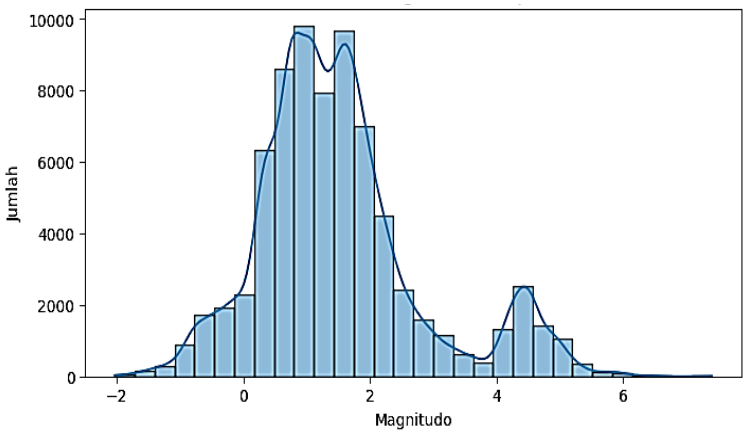
BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

3. Hasil dan Pembahasan

Proses analisis data gempa bumi dalam penelitian ini telah menghasilkan temuan penting terkait kinerja model prediksi magnitudo menggunakan pendekatan integrasi augmentasi data dan machine learning berbasis Random Forest Regressor(Picozzi & Iaccarino, 2021). Dalam tahap awal, data mentah dari sumber USGS (all_month.csv) berhasil diproses secara menyeluruh, termasuk pembersihan dari peristiwa non-seismik seperti *explosion* dan *quarry blast*, serta seleksi fitur kunci yang relevan untuk prediksi magnitudo. Setelah itu, teknik augmentasi data berbasis noise injection pada koordinat geografis dan kedalaman berhasil meningkatkan jumlah sampel sebesar 400% tanpa mengubah distribusi logis dari lokasi dan parameter fisik gempa bumi(Zhuang et al., 2023). Proses ini tidak hanya membantu menyeimbangkan data, tetapi juga memberikan variasi spasial yang lebih realistis, sehingga memperkuat kemampuan model dalam mengenali pola-pola kompleks di luar data latih.

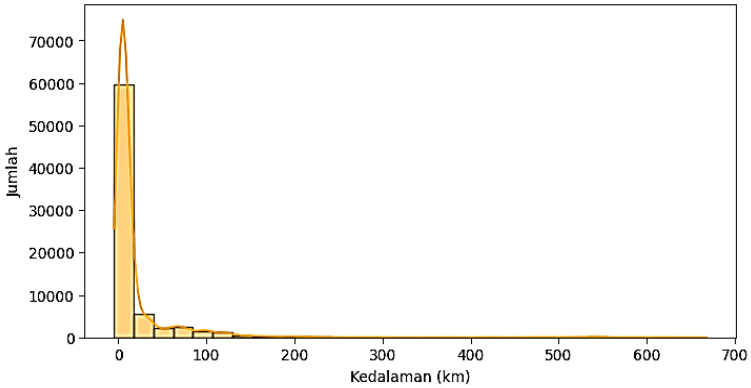
3.1. Visualisasi Dataset

Visualisasi distribusi magnitudo gempa bumi dalam Gambar 3 tersebut menunjukkan pola yang sangat khas dan relevan dengan fenomena seismik global. Grafik ini adalah histogram dengan kurva densitas superimpose, yang memberikan wawasan tentang bagaimana magnitudo gempa tersebar di dataset. Dari visualisasi ini dapat dilihat bahwa distribusi magnitudo memiliki karakteristik eksponensial atau log-normal, dimana jumlah gempa dengan magnitudo rendah (kurang dari 2) jauh lebih banyak dibandingkan dengan gempa besar (magnitudo > 4).



Gambar 3. Distribusi Magnitudo Gempa

Gambar 3 menunjukkan distribusi magnitudo gempa bumi dengan puncak histogram pada rentang 0–2, menandakan mayoritas peristiwa dalam dataset merupakan gempa kecil yang umumnya tidak terasa atau berdampak minimal. Kurva densitas memperlihatkan tren penurunan eksponensial seiring meningkatnya magnitudo, sesuai hukum Gutenberg-Richter yang menyatakan frekuensi gempa berkurang secara eksponensial dengan bertambahnya magnitudo. Jumlah gempa magnitudo 3–5 jauh lebih sedikit, sementara gempa besar (>5) sangat jarang terjadi. Pola ini menunjukkan ketidakseimbangan data, dengan dominasi gempa kecil. Kondisi ini menjadi tantangan prediksi, sehingga diperlukan teknik augmentasi untuk meningkatkan representasi gempa besar dalam pelatihan model machine learning.



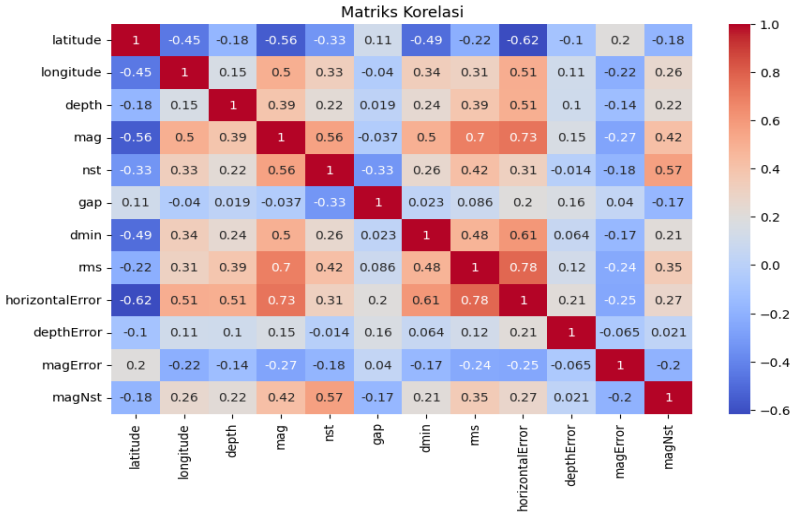
BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

Gambar 4. Distribusi Kedalaman Gempa

Gambar 4 menampilkan distribusi kedalaman gempa bumi dalam bentuk histogram dengan kurva densitas. Sebagian besar gempa terjadi pada kedalaman dangkal (0–10 km), sesuai dengan aktivitas tektonik intensif di zona subduksi atau batas lempeng. Frekuensi gempa menurun tajam seiring bertambahnya kedalaman, dengan jumlah minimal pada kedalaman >100 km. Outlier pada kedalaman >200 km kemungkinan terkait gempa intraplate atau aktivitas tektonik khusus, seperti di lempeng Pasifik atau India. Pola ini menegaskan bahwa mayoritas aktivitas seismik terkonsentrasi dekat permukaan, memberikan wawasan penting bagi analisis sumber dan dinamika gempa bumi global.

3.2. Korelasi atribut dataset

Analisis matriks korelasi yang disajikan pada Gambar 5 menunjukkan hubungan bervariasi antar fitur dengan magnitudo gempa. Fitur nst berkorelasi positif moderat (~0,56), depth positif sedang (~0,39), sedangkan gap (-0,33) dan dmin (-0,49) berkorelasi negatif. Artinya, lebih banyak stasiun dan gempa dangkal cenderung memiliki magnitudo lebih tinggi, sedangkan sudut stasiun besar dan jarak dekat ke stasiun sering terkait gempa kecil.



Gambar 5. Hasil Analisis Korelasi antar Atribut

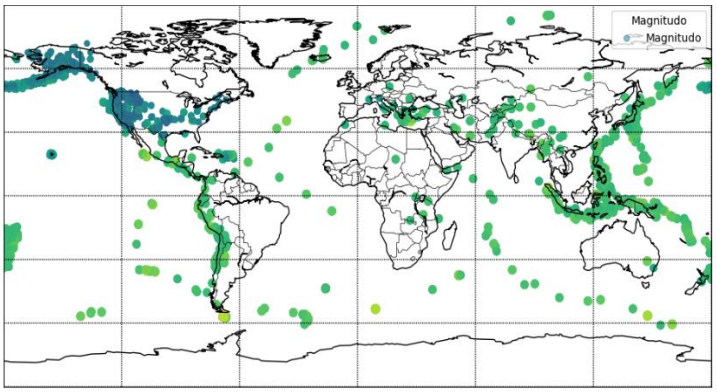
Gambar 5 menjelaskan tentang korelasi antar fitur yang memberikan informasi penting. Misalnya, horizontalError dan depthError memiliki korelasi positif kuat (sekitar 0.78), menunjukkan bahwa ketidakpastian dalam koordinat horizontal dan kedalaman umumnya berkorelasi satu sama lain. Ini relevan karena kesalahan dalam pengukuran spasial dan kedalaman sering kali bersamaan, terutama pada gempa dengan jumlah stasiun pencatat yang rendah.Secara keseluruhan, hasil analisis korelasi ini membantu memahami pola interkoneksi antar variabel dalam dataset, yang dapat digunakan untuk memilih fitur yang paling relevan dalam model prediksi magnitudo gempa bumi. Fitur dengan korelasi tinggi dengan magnitudo, seperti nst, gap, dan depth, kemungkinan besar akan menjadi input utama dalam model machine learning seperti Random Forest Regressor. Selain itu, pemahaman tentang korelasi antar fitur juga membantu menghindari masalah multicolinearity, sehingga model lebih stabil dan interpretasi lebih akurat.

3.3. Visualisasi Geospasial

Visualisasi peta lokasi gempa bumi pada Gambar 6 tersebut menunjukkan sebaran global dari peristiwa gempa berdasarkan magnitudo, dengan setiap titik mewakili lokasi suatu gempa. Peta ini menggunakan warna untuk merepresentasikan besarnya magnitudo: warna hijau tua mengindikasikan gempa dengan magnitudo lebih besar (gempa besar), sedangkan warna hijau terang menandakan gempa kecil hingga sedang. Dari visualisasi ini, dapat dilihat bahwa sebaran gempa bumi sangat heterogen di seluruh dunia, dengan konsentrasi yang lebih tinggi di wilayah-wilayah tertentu seperti Pasifik Ring of Fire, termasuk daerah sekitar Alaska, Jepang, Indonesia, dan Amerika Selatan. Wilayah lain seperti California, Meksiko, serta beberapa bagian Afrika juga menjadi zona aktivitas seismik yang signifikan. Di sisi lain, wilayah Eropa dan Afrika Tengah tampak memiliki frekuensi gempa yang lebih rendah dibandingkan dengan wilayah lempeng aktif

BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

seperti pada Gambar 6.



Gambar 6. Distribusi Lokasi Gempa Berdasarkan Magnitudo

Gambar 6 menjelaskan pola distribusi geologis global, di mana gempa cenderung terjadi di dekat batas-batas lempeng tektonik atau di wilayah dengan aktivitas vulkanik intensif. Misalnya, rantai gunung api di Asia Tenggara dan Amerika Selatan menunjukkan aktivitas seismik yang lebih sering dibandingkan dengan wilayah dataran stabil seperti Australia tengah atau Afrika Barat. Visualisasi ini tidak hanya memberikan wawasan tentang pola spasial gempa bumi tetapi juga membantu memahami hubungan antara lokasi gempa dan kondisi geologis di wilayah tersebut. Selain itu, penggunaan warna berbasis magnitudo memudahkan dalam membedakan intensitas gempa di berbagai wilayah, sehingga dapat digunakan sebagai alat penting dalam analisis risiko gempa bumi dan perencanaan mitigasi bencana.

3.3. Evaluasi Kinerja Model

Tabel perbandingan hasil evaluasi kinerja model Random Forest Regressor sebelum dan sesudah augmentasi data secara jelas menunjukkan peningkatan signifikan dalam akurasi dan stabilitas prediksi magnitudo gempa bumi setelah penerapan teknik augmentasi. Secara keseluruhan, semua metrik evaluasi menunjukkan penurunan yang signifikan: MAE turun dari 0,2467 menjadi 0,1046 (penurunan 57,6%), MSE berkurang dari 0,1225 ke 0,0349 (turun 71,5%), RMSE juga mengalami penurunan drastis dari 0,3499 ke 0,1868 (turun 46,6%), sementara R^2 meningkat dari 0,9493 menjadi 0,9817, menandakan bahwa model dengan augmentasi mampu menjelaskan varians target lebih baik. Peningkatan ini membuktikan bahwa augmentasi data berbasis noise injection pada fitur spasial dan kedalaman sangat efektif untuk mengatasi masalah ketidakseimbangan data dan meningkatkan kemampuan model dalam mempelajari pola-pola kompleks, terutama untuk kelas minoritas seperti gempa besar yang langka namun berdampak tinggi seperti yang disajikan pada Tabel 2.

Tabel 2. Evaluasi Model sebelum dan Sesudah Augmentasi Data

Matriks Evaluasi	**Sebelum Augmentasi Data**	**Sesudah Augmentasi Data**	**Deskripsi**
MAE (Mean Absolute Error)	0.2467	0.1046	↓ Berkurang sebesar57.6%
MSE (Mean Squared Error)	0.1225	0.0349	↓ Berkurang sebesar71.5%
RMSE (Root Mean Squared Error)	0.3499	0.1868	↓ Berkurang sebesar46.6%
R^2 (Coefficient of Determination)	0.9493	0.9817	↑ Meningkat sebesar3.4%

Tabel 2 menunjukkan bahwa augmentasi data secara signifikan meningkatkan kinerja model. Nilai MAE, MSE, dan RMSE menurun masing-masing sebesar 57,6%, 71,5%, dan 46,6%, menandakan berkurangnya kesalahan prediksi. Sementara itu, nilai R^2 meningkat 3,4%, mengindikasikan kemampuan model menjelaskan variasi data lebih baik. Secara keseluruhan, augmentasi data membuat prediksi lebih akurat dan model lebih andal.

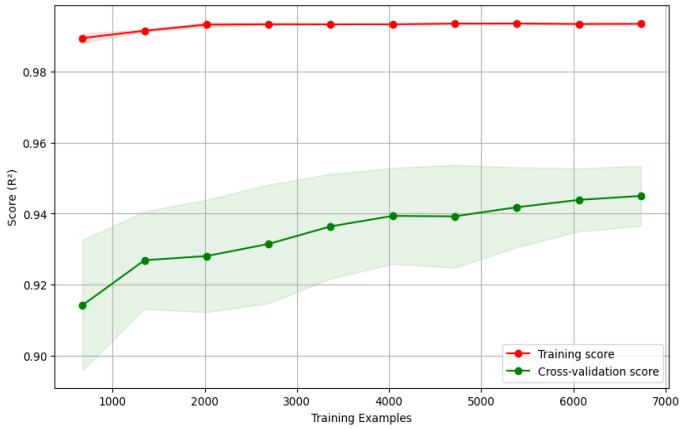
BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

3.3. Learning Curve Model

Learning curve digunakan untuk mengevaluasi kinerja model Random Forest Regressor dalam memprediksi magnitudo gempa bumi sebelum dan sesudah augmentasi data. Grafik ini menampilkan perubahan training loss (MSE) dan cross-validation loss (MSE) terhadap jumlah data pelatihan, sehingga terlihat kemampuan belajar model serta indikasi overfitting atau underfitting. Perbandingan sebelum dan sesudah augmentasi menunjukkan efektivitas teknik tersebut dalam meningkatkan generalisasi dan akurasi. Learning curve juga membantu mengidentifikasi titik optimal pelatihan, saat performa mulai stabil atau tambahan data tidak lagi signifikan, sekaligus menilai apakah model telah mencapai performa maksimal atau masih dapat ditingkatkan.

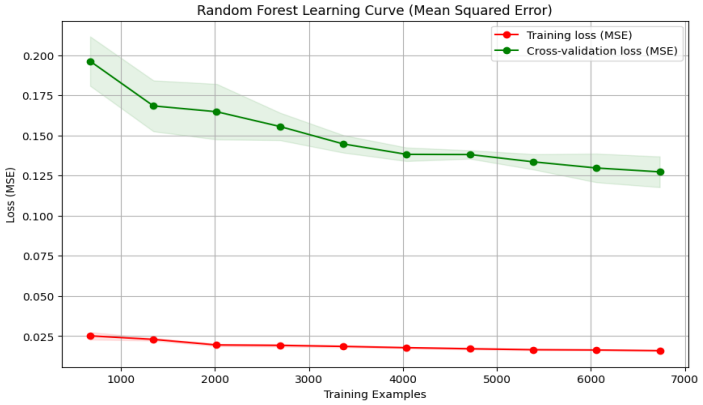
3.3.1 Learning Curve Sebelum Augmentasi Data

Grafik learning curve pada Gambar 7 menunjukkan perbandingan antara skor training score (dalam warna merah) dan cross-validation score (dalam warna hijau) terhadap jumlah contoh pelatihan. Dari visualisasi ini, dapat dilihat bahwa training score tetap stabil di tingkat yang sangat tinggi (sekitar 0.98–0.99), menunjukkan bahwa model mampu mempelajari pola dalam data latih dengan baik. Namun, cross-validation score, yang merepresentasikan kinerja model pada data uji independen, meningkat secara bertahap saat jumlah data pelatihan bertambah, meskipun masih berada di bawah training score. Ini mengindikasikan adanya sedikit overfitting, karena model lebih baik dalam memprediksi data yang digunakan untuk pelatihan dibandingkan data baru yang disajikan pada Gambar 7.



Gambar 7. Grafik Learning Curve(R^2 Score) sebelum Augmentasi Data.

Selain itu, area hijau transparan di sekitar garis cross-validation score menunjukkan variasi atau ketidakpastian dalam hasil validasi silang, yang semakin sempit seiring dengan penambahan jumlah data pelatihan. Hal ini menunjukkan bahwa dengan lebih banyak data, model menjadi lebih stabil dan konsisten dalam prediksi. Meskipun cross-validation score meningkat, jarak antara training score dan cross-validation score tetap cukup besar, menyiratkan bahwa model masih memiliki potensi untuk ditingkatkan agar lebih generalisasi, terutama jika ada lebih banyak data atau teknik seperti augmentasi data diterapkan untuk mengurangi overfitting.

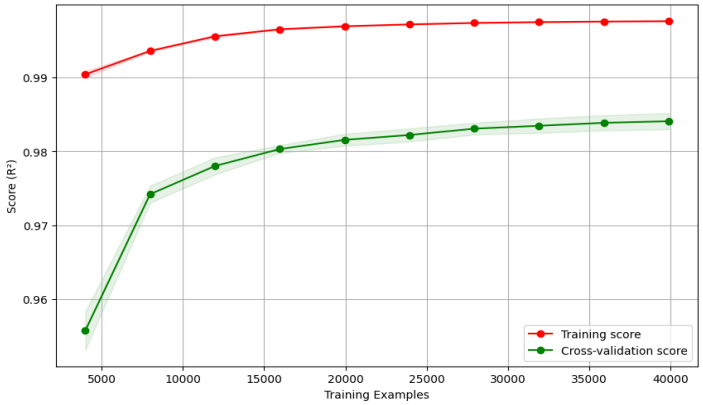


BHATARA: Jurnal Multidisiplin
Vol. 2 No. 3, 2025, pp. 97-108 ISSN: 3047-119

Gambar 8. Grafik Learning Curve(Mean Squared Error) sebelum Augmentasi Data

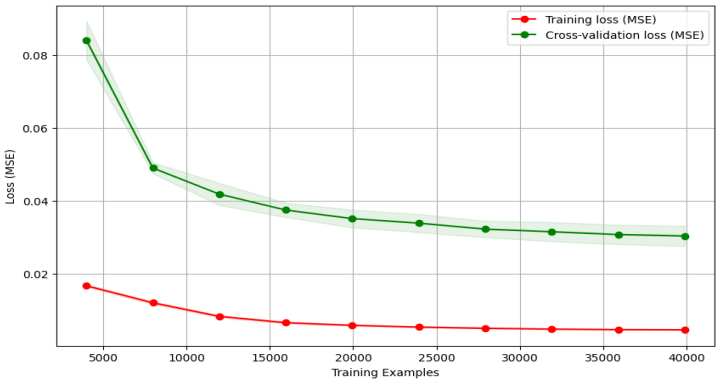
Gambar 8 menunjukkan learning curve Random Forest Regressor sebelum augmentasi data, membandingkan training loss (MSE, merah) dan cross-validation loss (MSE, hijau) terhadap jumlah data latih. Training loss stabil pada nilai rendah (~0,025), menandakan model mempelajari pola dengan baik. Cross-validation loss menurun seiring penambahan data, namun tetap lebih tinggi dari training loss, mengindikasikan sedikit overfitting. Area hijau transparan menunjukkan variasi validasi silang yang menyempit, menandakan peningkatan konsistensi. Perbedaan signifikan antara kedua kurva menunjukkan model masih dapat ditingkatkan generalisasinya, misalnya melalui penambahan data atau penerapan augmentasi untuk mengurangi overfitting.

3.3.2 Learning Curve Setelah Augmentasi Data



Gambar 9. Grafik Learning Curve(R^2 Score) sebelum Augmentasi Data

Gambar 9 menunjukkan learning curve Random Forest Regressor setelah augmentasi data, membandingkan training score (merah) dan cross-validation score (hijau). Hasilnya memperlihatkan peningkatan signifikan dibanding sebelum augmentasi, dengan training score stabil (~0,99) dan cross-validation score meningkat mendekati nilai training score, menandakan berkurangnya overfitting dan peningkatan generalisasi. Variasi pada validasi silang juga lebih kecil, menunjukkan stabilitas prediksi yang lebih baik. Jarak antara kedua kurva semakin sempit, mengindikasikan model lebih seimbang dan mampu menangani data baru dengan baik. Augmentasi data terbukti memperkuat kemampuan model dalam mempelajari pola kompleks, termasuk pada gempa besar yang jarang terjadi.



Gambar 10. Grafik Learning Curve(Mean Squared Error) Setelah Augmentasi Data

Selain itu, area hijau transparan pada Gambar 10 di sekitar garis cross-validation score menunjukkan variasi atau ketidakpastian dalam hasil validasi silang, yang semakin sempit seiring dengan penambahan jumlah data pelatihan. Hal ini menunjukkan bahwa dengan lebih banyak data, model menjadi lebih stabil dan konsisten dalam prediksi. Meskipun cross-validation score meningkat, jarak antara training score dan cross-validation score tetap cukup besar, menyiratkan bahwa model masih memiliki potensi untuk ditingkatkan agar lebih generalisasi, terutama jika ada lebih banyak data atau teknik seperti augmentasi data diterapkan untuk mengurangi overfitting.

4. Kesimpulan

Penelitian ini menunjukkan bahwa integrasi teknik augmentasi data dengan algoritma Random Forest Regressor merupakan pendekatan yang efektif dan menjanjikan untuk meningkatkan akurasi prediksi magnitudo gempa bumi. Analisis terhadap dataset USGS mengungkap distribusi magnitudo yang sangat tidak seimbang, dengan dominasi gempa kecil dan kelangkaan gempa besar, sehingga menjadi tantangan dalam membangun model prediksi yang robust. Penerapan augmentasi berbasis noise injection pada fitur spasial (latitude, longitude) dan kedalaman berhasil meningkatkan jumlah data latih hingga lima kali lipat tanpa mengurangi integritas geografis maupun fisik. Evaluasi kinerja model menunjukkan peningkatan signifikan: MAE menurun 57,6% (0,2467 → 0,1046), RMSE berkurang 46,6% (0,3499 → 0,1868), MSE turun 71,5% (0,1225 → 0,0349), dan R^2 meningkat dari 0,9493 menjadi 0,9817. Peningkatan ini membuktikan bahwa augmentasi tidak hanya mengurangi overfitting, tetapi juga memperkuat kemampuan model memprediksi gempa besar yang langka. Visualisasi geospasial mengonfirmasi kesesuaian sebaran gempa dengan zona lempeng aktif seperti Ring of Fire, mendukung validitas analisis. Secara keseluruhan, pendekatan ini memiliki potensi besar untuk diintegrasikan dalam sistem peringatan dini, strategi mitigasi bencana, dan pemetaan risiko seismik pada skala nasional maupun global.

Daftar Pustaka

1. Aden-Antoniów, F., Frank, W. B., & Seydoux, L. (2022). An Adaptable Random Forest Model for the Declustering of Earthquake Catalogs. *Journal of Geophysical Research: Solid Earth*, *127*(2). https://doi.org/10.1029/2021JB023254
2. Banna, M. H. Al, Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S. M. S., & Cho, G. H. (2020). Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges. *IEEE Access*, *8*, 192880–192923. https://doi.org/10.1109/ACCESS.2020.3029859
3. Bilal, M. A., Ji, Y., Wang, Y., Akhter, M. P., & Yaqub, M. (2022). An Early Warning System for Earthquake Prediction from Seismic Data Using Batch Normalized Graph Convolutional Neural Network with Attention Mechanism (BNGCNNATT). *Sensors*, *22*(17), 6482. https://doi.org/10.3390/s22176482
4. Chelidze, T., Kiria, T., Melikadze, G., Jimsheladze, T., & Kobzev, G. (2022). Earthquake Forecast as a Machine Learning Problem for Imbalanced Datasets: Example of Georgia, Caucasus. *Frontiers in Earth Science*, *10*. https://doi.org/10.3389/feart.2022.847808
5. D'Amico, E., Gabbolini, G., Bernardis, C., & Cremonesi, P. (2022). Analyzing and improving stability of matrix factorization for recommender systems. *Journal of Intelligent Information Systems*, *58*(2), 255–285. https://doi.org/10.1007/s10844-021-00686-1
6. Fornasari, S. F., Pazzi, V., & Costa, G. (2022). A Machine-Learning Approach for the Reconstruction of Ground-Shaking Fields in Real Time. *Bulletin of the Seismological Society of America*, *112*(5), 2642–2652. https://doi.org/10.1785/0120220034
7. Frith, K. H. (2023). ChatGPT: Disruptive Educational Technology. *Nursing Education Perspectives*, *44*(3), 198–199. https://doi.org/10.1097/01.NEP.0000000000001129
8. Ghimire, S., Guéguen, P., Giffard-Roisin, S., & Schorlemmer, D. (2022). Testing machine learning models for seismic damage prediction at a regional scale using building-damage dataset compiled after the 2015 Gorkha Nepal earthquake. *Earthquake Spectra*, *38*(4), 2970–2993. https://doi.org/10.1177/87552930221106495
9. Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, *117*, 47–58. https://doi.org/10.1016/j.future.2020.11.022
10. Jamjoom, M., & El Hindi, K. (2016). Partial instance reduction for noise elimination. *Pattern Recognition Letters*, *74*, 30–37. https://doi.org/10.1016/j.patrec.2016.01.021
11. Liu, J., Zhang, T., Gao, C., & Wang, P. (2023). Forecasting Earthquake Magnitude and Epicenter by Incorporating Spatiotemporal Priors Into Deep Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–13. https://doi.org/10.1109/TGRS.2023.3281784
12. Manral, G. S., & Chaudhary, A. (2023). Prediction of Earthquake Using Machine Learning Algorithms. *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, 1–5. https://doi.org/10.1109/ICIEM59379.2023.10166658
13. Mori, F., Mendicelli, A., Falcone, G., Acunzo, G., Spacagna, R. L., Naso, G., & Moscatelli, M. (2022). Ground motion prediction maps using seismic-microzonation data and machine learning. *Natural Hazards and Earth System Sciences*, *22*(3), 947–966. https://doi.org/10.5194/nhess-22-947-2022
14. N, M., & Thirukkumaran, R. (2020). Earthquake Prediction using Seismic Information. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *6*(3),

554–558. https://doi.org/10.32628/CSEIT2063107
15. Pasupulety, U., Abdullah Anees, A., Anmol, S., & Mohan, B. R. (2019). Predicting stock prices using ensemble learning and sentiment analysis. *Proceedings - IEEE 2nd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2019*, 215–222. https://doi.org/10.1109/AIKE.2019.00045
16. Picozzi, M., & Iaccarino, A. G. (2021). Forecasting the Preparatory Phase of Induced Earthquakes by Recurrent Neural Network. *Forecasting*, *3*(1), 17–36. https://doi.org/10.3390/forecast3010002
17. Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, *106*, 92–104. https://doi.org/10.1016/j.future.2020.01.005
18. Saleem, A. K., & Rashid, A. N. (2023). Applications of machine learning for earthquake prediction: A review. *Nucleation and Atmospheric Aerosols*, 030042. https://doi.org/10.1063/5.0119623
19. Saputro, I. W., & Sari, B. W. (2019). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Citec*, *6*(2460–4259), 1–11. https://doi.org/10.24076/citec.2019v6i1.178
20. Sharma, A., Ahuja, A., Devi, S., & Pasari, S. (2022). Use of Spatio-temporal Features for Earthquake Forecasting of imbalanced Data. *2022 International Conference on Intelligent Innovations in Engineering and Technology (ICIIET)*, 178–182. https://doi.org/10.1109/ICIIET55458.2022.9967687
21. Sugiyama, D., Tsuboi, S., & Yukutake, Y. (2021). Application of deep learning-based neural networks using theoretical seismograms as training data for locating earthquakes in the Hakone volcanic region, Japan. *Earth, Planets and Space*, *73*(1), 135. https://doi.org/10.1186/s40623-021-01461-w
22. Varshney, N., Kumar, G., Kumar, A., Pandey, S. K., Singh, T., & Singh, K. U. (2023). Machine Learning Based Algorithm for Earthquake Monitoring. *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, 264–270. https://doi.org/10.1109/CSNT57126.2023.10134640
23. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). Data Mining Practical Machine Learning Tools and Techniques Fourth Edition. In *Data Mining* (4th ed.). Morgan Kaufman. https://doi.org/10.1016/C2009-0-19715-5
24. Zhang, C., Zhang, Q., Wei, X., Guo, Z., Wang, Y., & Kim, S. (2023). Automatic first break picking with structured random forests. *Geophysical Prospecting*, *71*(8), 1473–1496. https://doi.org/10.1111/1365-2478.13390
25. Zhao, J. (2017). Combing semantic and prior polarity features for boosting twitter sentiment analysis using ensemble learning. *Proceedings - 2016 IEEE 1st International Conference on Data Science in Cyberspace, DSC 2016*, 709–714. https://doi.org/10.1109/DSC.2016.124
26. Zheng, Z., Wang, J., Shi, L., Zhao, S., Hou, J., Sun, L., & Dong, L. (2022). Generating phone-quality records to train machine learning models for smartphone-based earthquake early warning. *Journal of Seismology*, *26*(3), 439–454. https://doi.org/10.1007/s10950-021-10058-5
27. Zhuang, L., Fei, H., & Hu, P. (2023). Knowledge-enhanced event relation extraction via event ontology prompt. *Information Fusion*, *100*. https://doi.org/10.1016/j.inffus.2023.101919