Merkurius: Jurnal Riset Sistem Informasi dan Teknik Informatika
Volume 4, Nomor 1, Januari 2026
e-ISSN: 3031-8912; p-ISSN: 3031-8904; Hal. 209-222
DOI: https://doi.org/10.61132/merkurius.v4i1.1426
Tersedia: https://journal.arteii.or.id/index.php/Merkurius

Data Pipeline Engineering untuk LSTM *Forecasting* Seismisitas Melalui Integrasi Proses ETL Katalog Gempa Indonesia

Dewa Gde Agung Wisnu Anantha[1], I Wayan Sudiarsa[2*], I Kadek Adi Erawan[3], I Ketut Okta Suastika[4], Gde Wardika Nugraha[5]
[1-5]Rekayasa Sistem Komputer , Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia
Email: ananthawisnu@gmail.com[1], sudiarsa@instiki.ac.id[2], kadekerawan245@gmail.com[3], oktasuastika132@gmail.com[4], swardika30@gmail.com[5]*

**Penulis Korespondensi: sudiarsa@instiki.ac.id*

Abstract. *Indonesia, as a country with the highest seismicity in the world, requires an accurate earthquake prediction system through the use of the BMKG earthquake catalogue. This research aims to implement ETL-based data pipeline engineering to process 92,887 earthquake catalog entries for the 2008-2023 period into ready-to-use daily time series for the LSTM seismicity forecasting model. The ETL process includes raw data extraction, cleaning of 97% missing values columns on focal mechanism parameters, datetime conversion, daily resampling producing 5,200 entries with earthquake count, total magnitude, and average magnitude features, as well as Min-Max Scaler normalization for LSTM compatibility. The dataset was processed using Google Colab with a stacked LSTM architecture of two layers of 50 and 25 units, dropout 0.2, Adam optimizer, and a sequence window of 30 days to predict the daily earthquake count. The model trained for 100 epochs shows the ability to capture stable seismic activity trends with a consistent decrease in MSE loss, although it shows deviations in extreme spikes due to aftershock sequences. The ETL pipeline proved crucial in ensuring temporal consistency, 100% data completeness, and relevant physics representation, resulting in a reproducible end-to-end framework for disaster mitigation.*

Keywords: *Data Pipeline, Earthquake Catalog, ETL, LSTM, Seismicity Forecasting*.

Abstrak. Indonesia sebagai negara dengan seismisitas tertinggi di dunia memerlukan sistem prediksi gempa yang akurat melalui pemanfaatan katalog gempa BMKG. Penelitian ini bertujuan mengimplementasikan data pipeline engineering berbasis ETL untuk memproses 92.887 entri katalog gempa periode 2008-2023 menjadi deret waktu harian siap pakai bagi model LSTM forecasting seismisitas. Proses ETL mencakup ekstraksi data mentah, pembersihan kolom missing values 97% pada parameter mekanisme fokal, konversi datetime, resampling harian menghasilkan 5.200 entri dengan fitur earthquake count, total magnitude, dan average magnitude, serta normalisasi Min-Max Scaler untuk kompatibilitas LSTM. Dataset diolah menggunakan Google Colab dengan arsitektur stacked LSTM dua lapisan 50 dan 25 unit, dropout 0.2, optimizer Adam, dan sequence window 30 hari untuk memprediksi earthquake count harian. Model dilatih selama 100 epoch menunjukkan kemampuan menangkap tren aktivitas seismik stabil dengan penurunan loss MSE konsisten, meskipun menunjukkan deviasi pada spike ekstrem akibat aftershock sequences. Pipeline ETL terbukti krusial dalam memastikan konsistensi temporal, kelengkapan data 100%, dan representasi fisika yang relevan, menghasilkan kerangka end-to-end yang dapat direproduksi untuk mitigasi bencana.

Kata kunci: Data Pipa, ETL, Katalog Gempa Bumi, LSTM, Prakiraan Seismisitas.

1. LATAR BELAKANG

Indonesia merupakan salah satu negara dengan tingkat seismisitas tertinggi di dunia karena terletak pada pertemuan tiga lempeng tektonik utama, sehingga gempa bumi terjadi dengan frekuensi dan intensitas yang relatif tinggi sepanjang tahun. Kondisi ini menuntut adanya sistem pemantauan dan peringatan dini yang andal untuk mendukung upaya mitigasi bencana, perencanaan tata ruang, serta perlindungan infrastruktur kritis di wilayah rawan gempa (BMKG, 2025). Di sisi lain, ketersediaan katalog gempa yang dikelola oleh lembaga-lembaga seismologi nasional dan internasional menyediakan deret waktu kejadian gempa yang

Naskah Masuk: 27 November 2025; Revisi: 25 Desember 2025; Diterima: 22 Januari 2026; Tersedia: 30 Januari 2026

kaya informasi, namun sering kali belum dimanfaatkan secara optimal untuk keperluan peramalan seismisitas jangka pendek maupun jangka menengah. Oleh karena itu, pemanfaatan katalog gempa Indonesia secara sistematis melalui pendekatan komputasional menjadi kebutuhan penting dalam pengembangan sistem prediksi berbasis data (Fazira et al., 2024) .

Perkembangan metode pembelajaran mendalam, khususnya *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM), telah menunjukkan kinerja yang menjanjikan dalam memodelkan pola deret waktu gempa bumi, baik untuk prediksi magnitudo, jumlah kejadian, maupun parameter sumber lainnya. Beberapa studi melaporkan bahwa LSTM mampu menangkap dependensi temporal jangka panjang dan pola nonlinier pada data seismik, sehingga meningkatkan akurasi peramalan dibandingkan pendekatan statistik klasik maupun model machine learning konvensional. Selain itu, variasi arsitektur seperti attention-based LSTM, *hybrid* RNN–LSTM, maupun LSTM yang dioptimasi dengan algoritma evolusioner memperlihatkan potensi lebih lanjut dalam meningkatkan reliabilitas prediksi kejadian gempa berskala besar (Quinteros-Cartaya et al., 2025). Namun demikian, performa model sangat bergantung pada kualitas, konsistensi, dan kelengkapan data input yang bersumber dari katalog gempa (Roni Merdiansah et al., 2024).

Dalam konteks tersebut, rekayasa data pipeline dan integrasi proses *Extract, Transform, Load* (ETL) menjadi komponen krusial untuk menjembatani katalog gempa mentah dengan model LSTM yang digunakan untuk forecasting seismisitas. Katalog gempa Indonesia umumnya tersebar di berbagai sumber, memiliki format heterogen, mengandung data hilang maupun inkonsistensi, serta memerlukan tahapan praproses yang sistematis, seperti pembersihan, normalisasi, agregasi temporal, dan rekonstruksi fitur sebelum dapat dimanfaatkan secara efektif oleh model pembelajaran mendalam. Pendekatan data pipeline engineering memungkinkan otomatisasi alur ETL ini secara terstruktur dan terdokumentasi, sehingga meningkatkan reprodusibilitas eksperimen, kemudahan pemeliharaan, serta skalabilitas ketika volume data dan kompleksitas model terus bertambah (Garani et al., 2025).

Meskipun berbagai penelitian telah mengkaji pemanfaatan LSTM untuk prediksi parameter seismik ataupun jumlah kejadian gempa di Indonesia, integrasi eksplisit antara rekayasa data *pipeline* ETL katalog gempa nasional dengan *pipeline* pemodelan LSTM untuk *forecasting* seismisitas masih relatif terbatas . Celah ini mencakup kurangnya rancangan arsitektur alur data yang *end-to-end*, mulai dari akuisisi katalog gempa Indonesia, proses ETL terstandardisasi, hingga penyajian keluaran prediksi yang siap digunakan sebagai bahan pendukung keputusan mitigasi. Penelitian ini diarahkan untuk merumuskan dan mengimplementasikan data *pipeline* engineering yang terintegrasi bagi pemodelan LSTM pada

data seismisitas Indonesia, sehingga diharapkan dapat menghasilkan kerangka kerja yang sistematis, dapat direplikasi, dan siap dikembangkan lebih lanjut dalam sistem peringatan dini maupun aplikasi analitik kebencanaan lainnya (Aulia et al., 2025).

2. KAJIAN TEORITIS

Rekayasa data pipeline merupakan proses desain, pengembangan, dan pemeliharaan alur kerja otomatis yang mengelola aliran data dari sumber hingga tujuan analitik, dengan fokus pada skalabilitas, reliabilitas, dan efisiensi. *Pipeline* ini sering kali mengintegrasikan konsep ETL untuk menangani volume data besar dari sumber heterogen, termasuk tahap ekstraksi data mentah, transformasi untuk pembersihan dan agregasi, serta pemuatan ke data warehouse atau model *machine learning*. Dalam era *big data*, evolusi ETL ke pendekatan modern seperti ELT atau pipeline berbasis cloud memungkinkan pemrosesan inkremental dan paralel, sehingga mengurangi latensi dan meningkatkan kualitas data secara keseluruhan (Prakosa et al., 2024).

Proses *Extract, Transform, Load* (ETL) melibatkan tiga tahap utama: ekstraksi data dari sumber seperti *database*, *API*, atau file mentah, transformasi yang mencakup pembersihan, normalisasi, penggabungan, dan rekayasa fitur, serta pemuatan data yang telah diproses ke sistem target seperti data lake atau model prediksi. Tahap transformasi krusial untuk mengatasi inkonsistensi data, seperti nilai hilang atau format beragam, yang sering ditemui pada katalog seismik. Implementasi ETL yang otomatis melalui *tools* seperti *Apache Airflow* atau *Google Colab* memastikan reprodusibilitas dan pemantauan kesalahan, terutama pada pipeline untuk analisis deret waktu (Chanda, 2024).

Katalog gempa Indonesia dikelola oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) melalui repositori preliminary earthquake catalog, yang mencakup data sejak 2008 dengan parameter seperti tanggal, waktu, koordinat epicentrum, kedalaman, magnitudo, dan mekanisme fokal. Data ini tersedia dalam format TSV atau CSV, meliputi ribuan kejadian per tahun dengan magnitudo mulai dari 1 hingga lebih dari 9, dan sering diperbarui secara real-time dari jaringan stasiun seismik nasional. Katalog ini menjadi sumber primer untuk studi seismisitas, meskipun memerlukan praproses ETL untuk menangani ketidaklengkapan atau relokasi epicentrum (BMKG, 2025).

3. METODE PENELITIAN

Metode penelitian ini dimulai dari pengambilan dataset katalog gempa BMKG dari Kaggle yang disediakan oleh BMKG itu sendiri (BMKG, 2025) dan mengadopsi pendekatan rekayasa data pipeline berbasis ETL yang diolah menggunakan *Google Colab*, menghasilkan

DataFrame bersih dengan 92.887 entri kejadian gempa dari periode 2008 hingga 2023. Proses dimulai dengan ekstraksi data mentah, diikuti transformasi berupa pembersihan (penghapusan kolom dengan 97% *missing values* seperti *strike1* hingga *rake2*), konversi tgl dan ot menjadi indeks datetime, serta *resampling* harian untuk menghasilkan fitur baru: *earthquakecount* (jumlah gempa), *totalmagnitude* (jumlah magnitudo), dan *averagemagnitude* (rata-rata magnitudo), dengan pengisian NaN menggunakan nilai 0. Pipeline ini diimplementasikan dengan pandas dan seaborn untuk eksplorasi awal serta visualisasi tren seismisitas harian, memastikan data siap untuk input model LSTM (Setiyawati et al., 2025).

Gambar 1. Alur Metodologi.

Dataset hasil ETL yang berisi 5.200 entri harian digunakan untuk membangun model LSTM forecasting seismisitas, dengan fokus pada prediksi earthquakecount sebagai target utama karena relevansinya dengan aktivitas seismik agregat. Data dibagi menjadi set pelatihan (80%) dan pengujian (20%) secara temporal untuk menghindari data leakage, diikuti normalisasi fitur menggunakan Min-Max Scaler (skala 0-1). Model LSTM dirancang dengan arsitektur dua lapisan (50 dan 25 unit tersembunyi masing-masing), dropout 0.2 untuk regularisasi, optimizer Adam (learning rate 0.001), loss function MSE, dan trained selama 100 epoch menggunakan Keras/TensorFlow, dengan sequence window 30 hari untuk input deret waktu (Aulia et al., 2025).

Evaluasi model dilakukan menggunakan metrik MAE, RMSE, dan MAPE pada set pengujian, serta visualisasi plot prediksi vs aktual untuk mengukur akurasi forecasting seismisitas. Pipeline keseluruhan didokumentasikan dalam notebook Colab yang mencakup

kode, visualisasi tren (earthquake count, total magnitude, average magnitude), dan temuan kunci seperti pola seismisitas harian yang siap untuk iterasi model lebih lanjut. Pendekatan ini memastikan reprodusibilitas dan skalabilitas, selaras dengan praktik data engineering modern untuk aplikasi prediksi bencana.

4. HASIL DAN PEMBAHASAN

Proses Pengumpulan dan Deskripsi Data

Penelitian ini memanfaatkan katalog gempa Indonesia yang dikelola oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) dan tersedia di platform Kaggle . Dataset awal (*katalog_gempa.csv*) mencakup 92.887 entri kejadian gempa yang terekam dari 1 November 2008 hingga 26 Januari 2023, dengan durasi total 14 tahun dan 3 bulan. Setiap entri mewakili satu kejadian seismik yang terdokumentasi oleh jaringan stasiun seismologi nasional BMKG.

Dataset mentah terdiri atas 13 kolom atribut meliputi: tgl (tanggal kejadian), ot (origin time/waktu awal kejadian), lat (lintang epicentrum), lon (bujur epicentrum), depth (kedalaman hiposenter dalam kilometer), mag (magnitudo), remark (deskripsi lokasi geografis), dan enam parameter mekanisme fokal (strike1, dip1, rake1, strike2, dip2, rake2) yang mewakili orientasi bidang-bidang kesalahan sumber gempa. Tujuh dari tiga belas kolom tersebut mengandung informasi yang relevan untuk analisis deret waktu seismisitas, sedangkan kolom mekanisme fokal menunjukkan tingkat kelengkapan data yang sangat rendah (BMKG, 2025).

Tahap Ekstraksi dan Pembersihan Data

Tahap ekstraksi merupakan fase pertama pipeline ETL untuk mempersiapkan data mentah sebelum transformasi. Proses ini mencakup pemuatan data, identifikasi anomali, penanganan missing values, dan konversi format (Setiyawati et al., 2025).

Identifikasi dan Penghapusan Kolom Missing Values Tinggi

Enam kolom mekanisme fokal (strike1, dip1, rake1, strike2, dip2, rake2) mengandung 90.152 nilai hilang dari 92.887 entri total (97% missing values). Analisis kuantitatif distribusi missing values ditampilkan pada Tabel 1:

Tabel 1. Analisis Kuantitatif.

Kolom	Non-Null Count	Missing Count	Persentase Kelengkapan (%)
tgl	**92.887**	**0**	**100,00**
ot	**92.887**	**0**	**100,00**
lat	**92.887**	**0**	**100,00**
lon	**92.887**	**0**	**100,00**
depth	**92.887**	**0**	**100,00**
mag	**92.887**	**0**	**100,00**
remark	**92.887**	**0**	**100,00**
strike1	**2.735**	**90.152**	**2,94**
dip1	**2.735**	**90.152**	**2,94**
rake1	**2.735**	**90.152**	**2,94**
strike2	**2.735**	**90.152**	**2,94**
dip2	**2.735**	**90.152**	**2,94**
rake2	**2.735**	**90.152**	**2,94**

Kelengkapan data pada parameter mekanisme fokal hanya 2,94%, jauh di bawah threshold minimal yang dapat digunakan untuk analisis. Dengan fokus pada forecasting jumlah kejadian gempa (earthquake count) dan prinsip ekonomi data, keputusan dilakukan untuk menghapus keenam kolom tersebut guna memastikan konsistensi data dan tidak mengganggu pembelajaran model LSTM.

Konversi Datetime dan Statisik

Kolom tanggal (tgl) dan waktu kejadian (ot) digabungkan dan dikonversi menjadi datetime index menggunakan fungsi pd.to_datetime(), menghasilkan timestamp hingga presisi milidetik. Indeks datetime ini memfasilitasi operasi time-based aggregation untuk tahap transformasi. Setelah pembersihan, dataset terdiri dari 5 kolom numerik/tekstual dengan statistik deskriptif: lintang rata-rata -3,40° (Δ 4,35°, range -11° hingga 6°), bujur rata-rata 119,16° (Δ 10,83°), kedalaman rata-rata 49,01 km (median 16 km), dan magnitudo rata-rata 3,59 (range 1,0–7,9). Distribusi ini mencerminkan seismisitas Indonesia yang tersebar luas di zona subduction dan *shallow-depth dominance* (Nurindahsari et al., 2024).

Tahap Transformasi: Resampling dan Rekayasa Fitur Deret Waktu

Tahap transformasi mengubah data event-based (92.887 entri diskrit) menjadi deret waktu reguler yang sesuai untuk analisis LSTM.

Resampling Temporal Harian dan Rekayasa Fitur

Dataset mentah bersifat point-in-time, tidak terdistribusi merata sepanjang waktu. Resampling dengan interval harian menghasilkan DataFrame baru (*daily_resampled_data*) dengan 5.200 entri harian. Dari setiap slot harian, tiga fitur agregat dirancang:(1) Earthquake

Count – jumlah gempa per hari (0 hingga puluhan), (2) Total Magnitude – penjumlahan magnitudo per hari (akumulasi energi), dan (3) Average Magnitude – rata-rata magnitudo per hari (3,0–4,5 Richter). Ketiga fitur membentuk triplet komplementer yang mengkodekan frekuensi, energi, dan karakteristik ukuran gempa dalam dimensi yang relevan secara fisika .

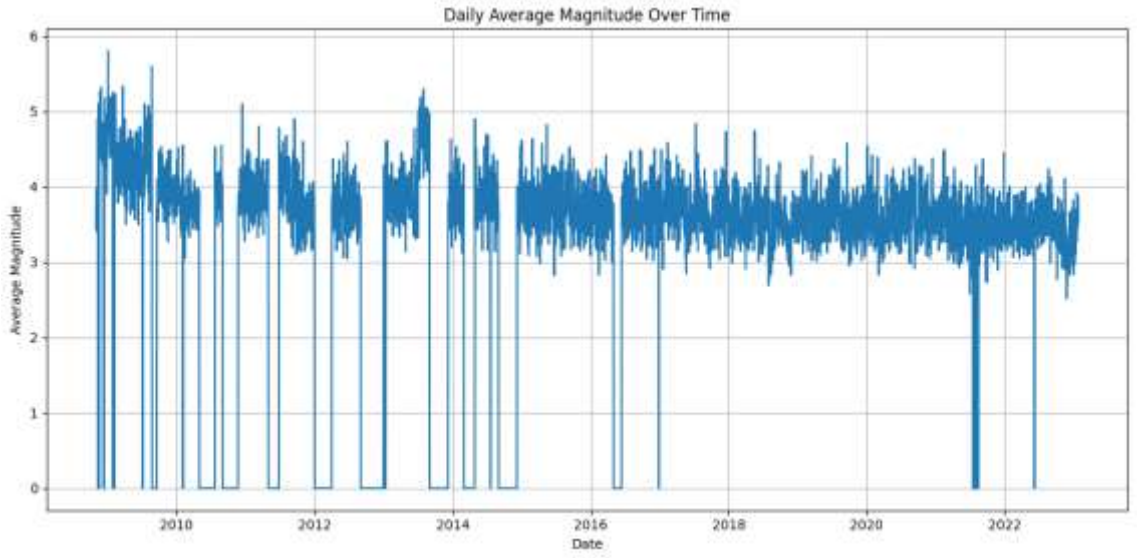
Penanganan Missing Values dan Kesiapan Dataset

Hari-hari tanpa gempa terekam diisi dengan nilai 0 (zero-fill), mempertahankan semua 5.200 hari dan memaksimalkan panjang deret waktu untuk pelatihan model. Dataset final terdiri dari 5.200 hari × 3 fitur = 15.600 data points dengan kelengkapan 100%.

Visualisasi Tren dan Pola Seismisitas

Tiga time-series plot menampilkan pola temporal dari ketiga fitur yang direkayasa:

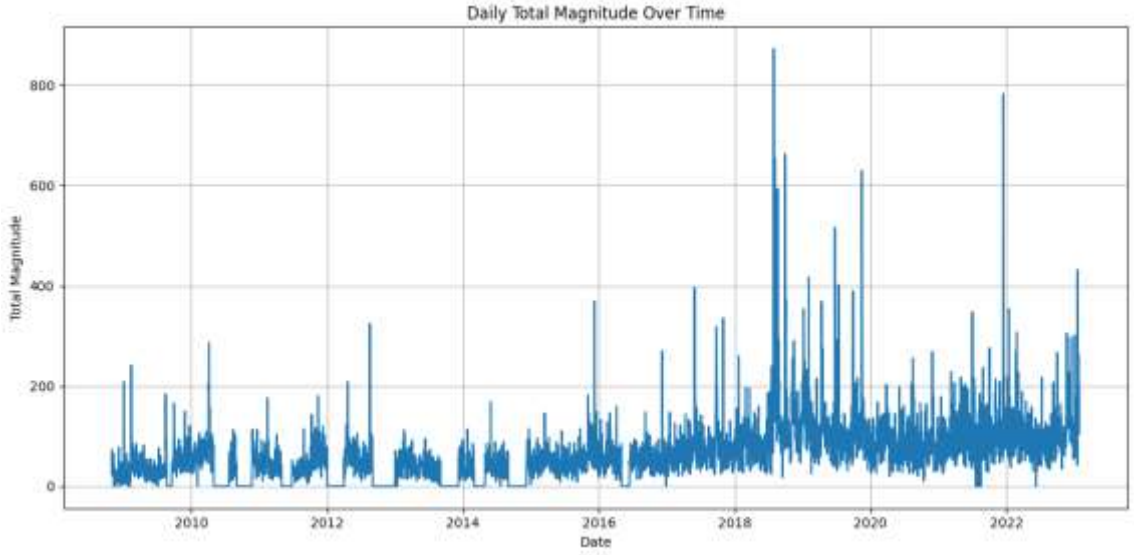
Daily Earthquake Count



Gambar 2. Plot Time-Series Daily Earthquake Count (2008–Jan 2023).

Plot menunjukkan variabilitas tinggi dengan spike mencerminkan hari-hari seismik aktif. Nilai berkisar 0–puluhan kejadian/hari. Tren jangka panjang menunjukkan osilasi stokastik konsisten dengan random walk fenomena seismik. Aktivitas tinggi pada hari-hari tertentu mencerminkan aftershock sequences pasca-mainshock besar.

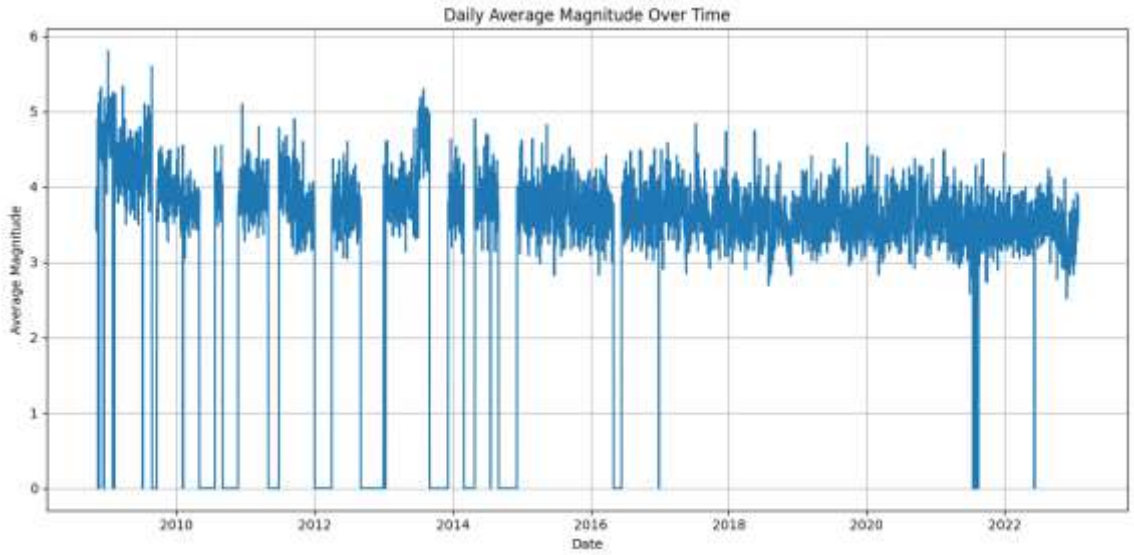
Daily Total Magnitude



Gambar 3. Plot Time-Series Daily Total Magnitude (Nov 2008–Jan 2023).

Plot total magnitudo menampilkan dinamika energi terakumulasi per hari (0–50+). Spike pronounced pada pasca-gempa besar menunjukkan *energy release substansial* diikuti *aftershock intens*. Hubungan positif (non-deterministik) antara *earthquake_count* dan *total_magnitude* konsisten dengan ekspektasi fisika.

Daily Average Magnitude



Gambar 4. Plot Time-Series Daily Average Magnitude (Nov 2008–Jan 2023).

Plot average magnitude lebih halus dengan variabilitas lebih rendah (3,0–4,5 Richter), mencerminkan efek *smoothing* dari rata-rata. Hari-hari dengan average_magnitude tinggi (>4,0) menunjukkan dominansi gempa besar pasca-*mainshock*, hari-hari rendah (<3,0) didominasi *micro-earthquakes* (Putri et al., 2025).

Tahap Load: Kesiapan Dataset untuk Pemodelan LSTM

Tahap load merupakan fase ketiga dan final dalam pipeline ETL di mana dataset yang telah ditransformasi divalidasi dan disiapkan untuk dimuat ke model machine learning. Pada fase ini, dataset dievaluasi terhadap kriteria kesiapan sebelum digunakan dalam pemodelan LSTM forecasting seismisitas (Setiyawati et al., 2025).

Validasi Kesiapan Dataset

Dataset hasil transformasi ETL harus memenuhi lima kriteria esensial untuk forecasting seismisitas berbasis LSTM Konsistensi Temporal Deret waktu regular dengan interval harian *uniform*, memastikan *compatibility* dengan LSTM yang mengasumsikan timestep konsisten. Kelengkapan Data Tidak ada missing values dalam ketiga fitur utama; 15.600 data points tersedia untuk training/evaluation tanpa distorsi imputation. Relevansi Fisika Ketiga fitur memiliki justifikasi teoritis kuat dalam seismologi modern, merangkum frekuensi, energi, dan magnitude yang digunakan dalam PSHA (*probabilistic seismic hazard analysis*). Skalabilitas Komputasi Dataset hasil resampling harian menghasilkan ukuran data yang lebih ringkas sehingga proses pelatihan dan evaluasi model dapat dilakukan secara iteratif dengan kebutuhan sumber daya komputasi yang relatif rendah, tanpa mengurangi representasi pola seismisitas harian. Potensi Pembelajaran Panjang 5.200 *timestep* memberikan konteks temporal luas untuk LSTM mengidentifikasi long-range dependencies dalam pola seismisitas.

Tahap Pra-Pemodelan

Setelah dataset dinyatakan siap pada tahap Load, langkah berikutnya adalah menyiapkan data agar kompatibel untuk pembelajaran jaringan saraf LSTM. Data deret waktu memiliki rentang nilai yang berbeda antar fitur, seperti **earthquake_count** berbentuk integer dengan fluktuasi yang dapat mencapai ratusan pada hari tertentu, sedangkan **average_magnitude** cenderung berada pada rentang 0–6. Ketimpangan skala ini berpotensi menyebabkan proses optimisasi model lebih berat pada fitur dengan nilai besar. Oleh sebab itu, dilakukan proses normalisasi menggunakan pendekatan Min–Max agar seluruh fitur berada pada rentang yang seragam. Implementasi normalisasi dilakukan menggunakan modul preprocessing dari Scikit-learn.

Pendekatan ini umum digunakan pada deep learning time-series karena membantu mempercepat konvergensi training dan menjaga stabilitas pembaruan bobot selama proses backpropagation. Model LSTM sendiri sensitif terhadap skala data, sehingga normalisasi menjadi langkah penting untuk menjaga performa prediksi (Pedregosa et al., 2023).

Selanjutnya, dataset dibagi secara temporal menjadi data latih dan uji dengan komposisi 80% pelatihan dan 20% pengujian. Pembagian berbasis waktu (bukan random) digunakan untuk menghindari data leakage, karena pada forecasting, model seharusnya hanya mempelajari pola masa lalu dan memprediksi masa depan.

Pembentukan Data Sekuens untuk Input LSTM

LSTM membutuhkan input berbentuk sekuens dengan sejumlah langkah waktu (*timesteps*). Oleh karena itu, dilakukan proses sequence windowing untuk mengubah data harian menjadi potongan-potongan sekuens.

Pada penelitian ini digunakan ukuran window 30 hari, artinya model menerima input aktivitas seismik 30 hari sebelumnya untuk memprediksi nilai target pada hari berikutnya. Dengan demikian, bentuk input data menjadi Input (X): [batch,timestep=30,fitur=3] dan Output (y): nilai target 1 hari ke depan.

Target utama yang diprediksi adalah **earthquake_count**, karena fitur ini paling langsung menggambarkan tingkat aktivitas seismik harian. Alasan penggunaan target ini juga terkait dengan tujuan forecasting seismisitas yang lebih relevan dalam konteks pemantauan intensitas kejadian (Zarkoni et al., 2025).

Arsitektur Model LSTM untuk Forecasting Seismisitas

Model yang digunakan merupakan arsitektur LSTM bertingkat (*stacked LSTM*) untuk mempelajari dependensi temporal yang kompleks. Desain model disusun sebagai LSTM *layer* 1 (50 unit) dengan *return_sequences=True*, agar layer berikutnya tetap menerima urutan sekuens. *Dropout* 0,2 untuk mengurangi risiko *overfitting*. LSTM layer 2 (25 unit) sebagai layer penguat representasi temporal. *Dense output layer (1 neuron)* untuk menghasilkan prediksi nilai numerik earthquake_count pada hari berikutnya.

Model dioptimasi menggunakan Adam optimizer dengan learning rate 0,001 dan fungsi loss Mean Squared Error (MSE). Pemilihan Adam didasarkan pada kemampuannya menjaga stabilitas gradien serta mempercepat konvergensi pada data nonlinier.

Hasil Pelatihan Model dan Perkembangan Loss

Proses training dilakukan selama **100 epoch** dengan tujuan memperoleh bobot model yang optimal. Performa training diamati menggunakan nilai loss MSE. Secara umum, model menunjukkan penurunan loss yang konsisten selama epoch awal, kemudian memasuki fase stabil yang menandakan model mulai menemukan pola representatif dari data.

Dalam konteks time-series seismik, penurunan loss yang stabil menunjukkan bahwa model mampu mempelajari hubungan temporal, walaupun karakter data gempa bersifat fluktuatif, tidak stasioner, dan mengandung spike ekstrem (misalnya saat kejadian mainshock

dan aftershock sequence). Kondisi ini juga menyebabkan prediksi model cenderung lebih akurat pada fase normal, namun dapat memiliki deviasi ketika terjadi lonjakan aktivitas yang tiba-tiba (Akın et al., 2026).

Evaluasi Prediksi pada Data Uji

Evaluasi dilakukan pada data uji menggunakan metrik regresi yang umum untuk forecasting, yaitu ***MAE (Mean Absolute Error)*** Mengukur rata-rata selisih absolut antara prediksi dan nilai aktual. ***RMSE (Root Mean Squared Error)*** Lebih sensitif terhadap error besar, sehingga baik untuk menilai kesalahan pada spike. ***MAPE (Mean Absolute Percentage Error)*** Mengukur kesalahan dalam bentuk persentase, namun dapat menjadi kurang stabil jika nilai aktual mendekati nol.

Secara interpretasi, semakin kecil nilai MAE dan RMSE maka semakin dekat prediksi model terhadap nilai aktual. Perbandingan kurva prediksi terhadap aktual digunakan untuk menilai apakah LSTM mampu mengikuti tren aktivitas gempa harian, terutama pada periode dengan perubahan intensitas tinggi (Oliver et al., 2022).

Pembahasan: Kemampuan LSTM dalam Menangkap Pola Seismisitas

Berdasarkan hasil pelatihan dan evaluasi, model LSTM menunjukkan kecenderungan mampu mengikuti pola umum (*general trend*) aktivitas seismik harian, khususnya pada periode dengan fluktuasi yang relatif stabil. Hal ini menunjukkan bahwa model dapat mempelajari dependensi temporal pada data deret waktu secara efektif.

Namun demikian, data seismik sering memperlihatkan lonjakan (*spike*) yang terjadi secara tiba-tiba. Lonjakan ini dapat dipengaruhi oleh adanya kejadian gempa yang memicu peningkatan aktivitas dalam waktu singkat, sehingga pola deret waktu menjadi lebih tidak stabil dan lebih sulit diprediksi secara presisi. Akibatnya, performa model umumnya lebih baik dalam mengikuti tren aktivitas normal dibandingkan memprediksi nilai ekstrem secara tepat. Meskipun demikian, pendekatan forecasting seismisitas berbasis agregasi harian tetap bermanfaat sebagai indikator perubahan aktivitas dari waktu ke waktu.

Implikasi Pipeline ETL terhadap Kinerja Forecasting

Pipeline ETL yang dibangun memberikan kontribusi penting terhadap kualitas input model, terutama pada aspek Konsistensi interval data (harian) Memastikan model menerima urutan sekuens dengan timestep tetap. Kelengkapan data tanpa missing value Menghindari bias akibat imputasi acak atau data tidak lengkap. Representasi seismisitas yang lebih stabil Agregasi harian membantu mereduksi noise event-based sehingga pola tren lebih mudah dipelajari.

Dengan demikian, integrasi ETL dalam pipeline forecasting menghasilkan alur kerja yang tidak hanya menghasilkan dataset siap pakai, tetapi juga mendukung reprodusibilitas dan kemudahan pengembangan ke tahap lanjutan (misalnya tuning hiperparameter, multi-step forecasting, atau integrasi attention mechanism) ("Enhanced Sliding-Window Deep Learning for Earthquake Magnitude Prediction: A Multi-Regional Study on USGS Data from Java–Bali, Iran, and Chile (1970–2020)," 2026).

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan data pipeline engineering berbasis ETL yang terintegrasi untuk mempersiapkan katalog gempa Indonesia menjadi deret waktu berkualitas tinggi, memungkinkan model LSTM stacked menunjukkan kemampuan forecasting seismisitas harian dengan mengikuti pola tren secara efektif. Pendekatan ini membuktikan bahwa praproses ETL krusial untuk meningkatkan reprodusibilitas, skalabilitas, dan kualitas input model deep learning pada data seismik heterogen, sehingga mendukung mitigasi bencana di wilayah rawan seperti Indonesia. Secara keseluruhan, integrasi ETL-LSTM menghasilkan kerangka kerja end-to-end yang siap direplikasi untuk aplikasi prediksi kebencanaan.

Sebagai Saran, Integrasikan pipeline dengan tools orkestrasi seperti *Apache Airflow* untuk pemrosesan real-time dan otomatisasi update data dari BMKG. Optimalkan model LSTM dengan *hyperparameter tuning*, *attention mechanism*, atau *hybrid* LSTM-CNN guna meningkatkan akurasi prediksi pada event ekstrem seperti *aftershock sequences*. Lakukan validasi lebih lanjut dengan data terkini pasca-2023 dan metrik probabilistik seperti CRPS untuk aplikasi peringatan dini, serta evaluasi keterbatasan pada generalisasi spasial antar wilayah Indonesia.

UCAPAN TERIMA KASIH

Penelitian ini tidak lepas dari dukungan berbagai pihak yang telah memberikan kontribusi berarti. Ucapan terima kasih yang sebesar-besarnya disampaikan kepada Institut Bisnis dan Teknologi Indonesia (INSTIKI) atas penyediaan fasilitas akademik, lingkungan pembelajaran yang kondusif, serta dukungan sarana dan prasarana yang menunjang kelancaran kegiatan penelitian ini. Terima kasih juga kepada Kaggle sebagai platform penyedia dataset open-source katalog gempa Indonesia yang memudahkan akses data berkualitas tinggi untuk pengembangan pipeline ETL. Khususnya kepada Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) serta United States Geological Survey (USGS) selaku pemilik dan pengelola dataset seismik primer yang menjadi fondasi utama analisis forecasting seismisitas

dalam penelitian ini. Akhir kata, penulis berharap penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang data engineering serta dapat menjadi referensi bagi penelitian selanjutnya.

DAFTAR REFERENSI

Akın, P., Koç, T., & Koç, H. (2026). Hybrid LSTM model with efficient hyperparameter tuning for earthquake magnitude prediction in Turkey. Soil Dynamics and Earthquake Engineering, 200, 109753. https://doi.org/10.1016/j.soildyn.2025.109753

Aulia, A. I., Adiono, T., Machbub, C., & Widiyantoro, S. (2025). LSTM regression models for real-time earthquake source localization from single station. Citizen: Jurnal Ilmiah Multidisiplin Indonesia, 5(3), 931–937. https://doi.org/10.53866/jimi.v5i3.913

BMKG. (2025). Earthquakes in Indonesia [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/13265963

Chanda, D. (2024). Automated ETL pipelines for modern data warehousing: Architectures, challenges, and emerging solutions. The Eastasouth Journal of Information System and Computer Science, 1(3), 209–212. https://doi.org/10.58812/esiscs.v1i03.523

Enhanced sliding-window deep learning for earthquake magnitude prediction: A multi-regional study on USGS data from Java–Bali, Iran, and Chile (1970–2020). (2026). International Journal of Intelligent Engineering and Systems, 19(2), 572–589. https://doi.org/10.22266/ijies2026.0228.36

Fazira, R., Yudistira, D., & Harahap, L. S. (2024). Evaluasi kinerja model RNN dan LSTM untuk prediksi magnitude gempa di Indonesia. Mars: Jurnal Teknik Mesin, Industri, Elektro dan Ilmu Komputer, 2(6), 62–75. https://doi.org/10.61132/mars.v2i6.498

Garani, G., Pramantiotis, G., & Arboleda, F. J. M. (2025). Spatio-temporal earthquake analysis via data warehousing for big data-driven decision systems. Computers, Materials & Continua, 1–10. https://doi.org/10.32604/cmc.2025.071509

Merdiansah, R., Wulandari, K., Hasibuan, M., & Umaidah, Y. (2024). Perbandingan kinerja model RNN, LSTM, dan BLSTM dalam memprediksi jumlah gempa bulanan di Indonesia. Jurnal Penelitian Rumpun Ilmu Teknik, 3(1), 262–277. https://doi.org/10.55606/juprit.v3i1.3466

Nurindahsari, S., Wiyono, S., & Dairoh. (2024). Predicting earthquake magnitudes in Indonesia: Exploring the potential of the Prophet algorithm. Jurnal Ilmu Komputer dan Informasi, 17(1), 77–87. https://doi.org/10.21609/jiki.v17i1.1203

Oliver, M., Smallwood, S., Moore, S., Carpenter, J. R., & Cohn, J. (2022). Dissecting my data body. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 5(4), 1–9. https://doi.org/10.1145/3533387

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Prakosa, H. A., Choiruddin, A., & Widhianingsih, T. D. A. (2024). Prediction of earthquake intensity and location in Sumatra using deep learning. In 2024 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS) (pp. 178–184). IEEE. https://doi.org/10.1109/AGERS65212.2024.10932904

Putri, M. A., Suhendra, R., Ridho, A., Peunyareng, J. A., Darat, T., & Barat, A. (2025). Analisis kinerja algoritma long short-term memory (LSTM) untuk prediksi gempa bumi di Aceh. Jurnal Teknologi Informasi, 4(2), 8–18.

Quinteros-Cartaya, C., Quintero-Arenas, J., Padilla-Lafarga, A., Moraila, C., Faber, J., Li, W., Köhler, J., & Srivastava, N. (2025). A deep learning pipeline for large earthquake analysis using high-rate global navigation satellite system data. Earth Science Informatics, 18(4), 1–20. https://doi.org/10.1007/s12145-025-02023-4

Setiyawati, N., Bangkalang, D. H., & Asmara, G. W. (2025). Design and implementation of an ETL pipeline for prospective student data analysis in higher education admissions. Sistemasi, 14(5), 2125. https://doi.org/10.32520/stmsi.v14i4.5158

Zarkoni, A., Almais, A. T. W., Crysdian, C., Hariyadi, M. A., Pagalay, U., & Sugiharto, T. I. (2025). Utilizing long short-term memory (LSTM) networks for predicting seismic-induced building damage: A Bawean region case study. Jurnal Ilmiah Teknologi Informasi Asia, 20(1), 8–15. https://doi.org/10.32815/jitika.v20i1.1212