

Forecasting Curah Hujan di Bojonegoro Menggunakan Model CNN-CA-LSTM dan Metode *Harmonic Dataset Distillation*

Kurniawan Indra Jaya^{a,1*}, Andi Mulyanti Suhartini^{a,2}, Nurul Jariyatin^{a,3}, Devi Septiani^{a,4}

^a Program Studi Rekayasa Perangkat Lunak, Fakultas Sains dan Teknologi, Institut Sain Teknologi Kesehatan Insan Cendekia Husada Bojonegoro, Bojonegoro, Indonesia

e-mail: ¹kurniawanindrajaya12317@gmail.com, ²andimulyanti.istekicsada@gmail.com, ³nj.icsada@gmail.com, ⁴deviseptiani569@gmail.com

*Corresponding author

Submitted Date: March 30th, 2026
Revised Date: May 25th, 2026

Reviewed Date: April 10th, 2026
Accepted Date: June 15th, 2026

Abstract

Rainfall forecasting in Bojonegoro region of Indonesia presents significant challenge due seasonal weather fluctuations and challenges of rainless days, which result in large number of zero data points or zero inflation. This study utilizes historical rainfall datasets in Bojonegoro since 1960, aggregated monthly with total 797 data, to explore deep learning architectures with Attention Module mechanisms and data optimization techniques to optimize model evaluation results. This study explores the CNN-LSTM model with modified Attention Module in the form of Coordinate Attention (CA) and evaluates it on baseline model. This study employs Harmonic Dataset Distillation (HDD) method, a novelty of this study, to improve computational efficiency through frequency domain data compression using the HDD method, and to maintain long-term seasonal patterns. The results of comparative analysis on CNN-LSTM with feature extraction model CNN-CA-LSTM Backbone model and the Basic Model can reduce training computation time by up to 73.66%. The Model CNN-CA-LSTM maintained good performance values in the Baseline Model, got an R^2 of 0.78 with an RMSE result of 68.38, with the CNN-CA-LSTM Proposed Model with CNN-CA-LSTM Backbone got an R^2 of 0.75 and an RMSE of 72.24.

Keywords: *harmonic dataset distillation; attention module; deep learning; forecasting*

Abstrak

Forecasting curah hujan di daerah Bojonegoro Indonesia menjadi suatu tantangan yang signifikan karena terdapat fluktuasi musiman cuaca serta tantangan pada hari tanpa hujan yang menyebabkan banyaknya data nol atau inflasi nol. Penelitian menggunakan dataset historis curah hujan di Bojonegoro sejak tahun 1960 diagregasi setiap bulan dengan total data 797 untuk mengeksplorasi tentang arsitektur deep learning dengan mekanisme Attention Module serta teknik optimasi pada data untuk mengoptimalkan hasil dari evaluasi model. Penelitian ini mengeksplorasi pada model CNN-LSTM dengan modifikasi Attention Module yang berupa Coordinate Attention (CA) serta mengevaluasinya pada model dasar. Penelitian ini menggunakan Pendekatan Metode Harmonic Dataset Distillation (HDD) yang merupakan kebaruan dari penelitian ini yang digunakan untuk meningkatkan efisiensi pada komputasi melalui kompresi data domain di frekuensi menggunakan metode HDD ini, serta mampu mempertahankan pola musiman jangka panjang. Hasil Analisis Komparatif pada model CNN-LSTM dengan model ekstraksi fitur atau Backbone CNN-CA-LSTM dan Model Dasar mampu memangkas waktu komputasi pelatihan hingga 73.66%. Model CNN-CA-LSTM pada Metode Dasar dan Metode Usulan mempertahankan nilai performa dengan baik pada Model Dasar yaitu R^2 sebesar 0.78 dengan hasil RMSE 68.38, dengan Model Usulan dengan Backbone CNN-CA-LSTM yaitu R^2 sebesar 0.75 dan RMSE sebesar 72.24.

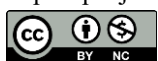
Kata Kunci: *harmonic dataset distillation; attention module; deep learning; forecasting*

1. Pendahuluan

Bojonegoro adalah salah satu wilayah di Jawa Timur yang mempunyai wilayah agraris yang cukup besar. Wilayah agraris ini juga dipengaruhi oleh banyaknya air pada suatu wilayah, seperti saat kekurangan air maka tempat ini akan mengalami kekeringan dan jika kebanyakan air maka akan mengakibatkan banjir dan itu dapat menimbulkan dampak seperti gagal panen jika terjadi kekeringan. Fluktuasi pada tingkat air yang disebabkan oleh perubahan iklim yang berada di Bojonegoro ini dapat menimbulkan penurunan tingkat kualitas hidup terutama akan berdampak pada masyarakat yang berprofesi sebagai petani yang cenderung kurang menyadari dari potensi perubahan iklim, serta banjir terus menerus dan dapat menyebabkan kerugian dan kehilangan modal [1]. Penelitian ini menggunakan metode analisis *Time series* untuk menyelesaikan permasalahannya [2]. Data cuaca seperti presipitasi curah hujan, memiliki karakteristik yang sangat fluktuatif, mulai dari karakter data di musim kemarau yang memiliki banyak nilai nol atau *Zero Inflated Data* dan kenaikan data fluktuasi data yang ekstrem saat musim hujan. Serta, saat pelatihan pada model kecerdasan buatan dapat memakan waktu yang cukup lama, sehingga terdapat permasalahan juga pada efisiensi pada tingkat komputasinya.

Pada penelitian sebelumnya, penelitian-penelitian dengan topik analisis *Time series* sudah banyak diterapkan seperti untuk prediksi atau *forecasting* pada presipitasi curah hujan [3], [4], [5]. Serta, dalam beberapa artikel tercatat bahwa lama pelatihan pada model kecerdasan buatan dapat mencapai puluhan menit untuk sekali pelatihan [6]. Penelitian sebelumnya tentang analisis *time series* pernah dilakukan oleh Indra Jaya [7], penelitian ini menggunakan data harga indeks saham Indonesia (IHSG), hasil penelitian tersebut menunjukkan model LSTM memiliki tingkat evaluasi yaitu nilai RMSE sebesar 93.68 dan R^2 sebesar 0.73, penelitian ini memiliki kelemahan, salah satunya yaitu tentang efisiensi waktu saat pelatihan dan komputasi karena model kecerdasan buatan memerlukan seluruh dari baris data asli pada pelatihannya. Terdapat penelitian oleh Wijayanto [8], untuk prediksi curah hujan untuk identifikasi potensi bencana banjir yang terdapat pada Kota Semarang dengan 1 parameter yaitu presipitasi curah hujan di Kota Semarang, penelitian ini menggunakan model LSTM untuk analisis *Time series*, didapatkan hasil MSE terbaik yaitu 19.11 dan RMSE terbaik yaitu 4.37, penelitian ini kekurangan yaitu pada Model yang digunakan belum terlalu kuat untuk menangkap pola spasial-temporal pada data *Time series*, yaitu fitur spasial merujuk pada ekstraksi fitur di karakteristik pola lokal dalam runtutan data dengan lapisan CNN dan fitur temporal merujuk pada ekstraksi karakteristik pada tren sekuensial dengan lapisan LSTM serta permasalahan waktu komputasi karena model kecerdasan buatan memerlukan seluruh dari baris data yang digunakan pada pelatihannya dan fluktuasi pada data presipitasi curah hujan yang tinggi. Berdasarkan penyajian penelitian sebelumnya, analisis *time series* sudah banyak menggunakan metode dari kecerdasan buatan seperti teknik *machine learning* maupun *deep learning* [9]. Tetapi, terdapat beberapa kekurangan seperti pada model CNN-LSTM pada saat pelatihan sering kehilangan fitur detail pada data yang memiliki tingkat fluktuasi data yang tinggi seperti pada data presipitasi curah hujan. Pelatihan dengan metode kecerdasan buatan pada tipe *deep learning* memiliki karakteristik yaitu membutuhkan data yang besar, karakteristik data pada *time series* presipitasi curah hujan yang memiliki tingkat fluktuasi pada parameter data yang tinggi, dan pelatihan model kecerdasan buatan yang kompleks ini pada sumber komputasi yang terbatas dapat menjadi hambatan.

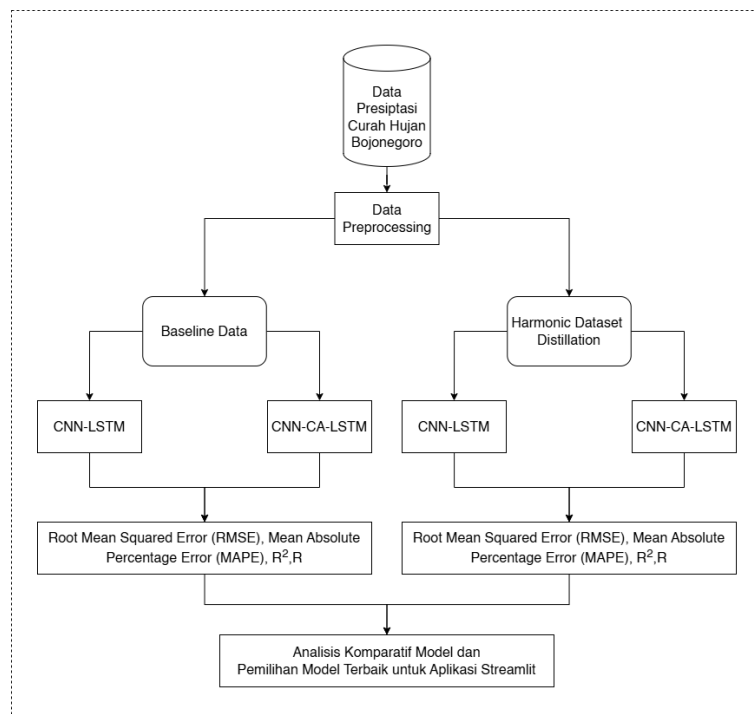
Penelitian sebelumnya hanya berfokus pada menggunakan model CNN-LSTM, LSTM, atau CNN-CA-LSTM untuk model prediksi *time series*, serta hanya berfokus pada penggunaan *dataset* asli dengan tahapan *preprocessing* data seperti penggunaan teknik normalisasi atau teknik augmentasi data [7], [8]. Penelitian ini mengeksplorasi pada model kecerdasan buatan untuk menjawab permasalahan pada fluktuasi data curah hujan yang terjadi setiap tahun khususnya pada penelitian ini menggunakan data di Bojonegoro, dengan mengusulkan model CNN-LSTM, dengan modifikasi penambahan modul *Attention* serta menggunakan Metode *Dataset Distillation* yaitu *Harmonic Dataset Distillation* (HDD) yang mempunyai kemampuan untuk menyaring *noise* fluktuasi pada presipitasi curah hujan yang tinggi serta untuk menyintesis data menjadi kumpulan data yang lebih kecil yang telah disintesis tetapi dengan hasil evaluasi pada kinerja model tersebut hampir sebanding dengan *dataset* asli [10]. Modul *Attention* yang digunakan yaitu *Coordinate Attention* (CA) dapat membantu untuk memperkuat serta pada model kecerdasan buatan lebih bisa menangkap informasi arah dan posisi dengan lebih baik [11], [12]. Penelitian ini, dievaluasi menggunakan matriks evaluasi yaitu *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), *Correlation Coefficient Pearson* (R), *Correlation Determination* (R^2), serta waktu pelatihan.



Penelitian ini bertujuan untuk membandingkan performa model dengan analisis komparatif dengan metode evaluasi yang digunakan pada tahapan baseline dengan menggunakan dataset asli dan tahapan usulan, yaitu menggunakan metode Harmonic Dataset Distillation (HDD) untuk mendapatkan dataset sintetis dengan meniru pola dataset asli, dengan menggunakan metode dasar yaitu CNN-LSTM dan CNN-CA-LSTM, serta mengeksplorasi model usulan dengan model backbone atau ekstraksi fitur menggunakan CNN-LSTM dan CNN-CA-LSTM. Kontribusi utama pada penelitian ini yaitu implementasi dari Metode HDD pada analisis data *time series* dapat meningkatkan efisiensi komputasi pada pelatihan Model dengan hasil evaluasi model memiliki performa yang mirip dengan *dataset* asli. *Dataset* sintetis dari HDD ini memiliki karakteristik yang hampir mirip dengan *dataset* asli namun memiliki ukuran yang lebih ringkas.

2. Metodologi Penelitian

Penelitian ini merupakan penelitian kuantitatif eksploratif pada topik analisis *Time series*. Penelitian ini mengeksplorasi pada Model Kecerdasan Buatan yaitu CNN-LSTM dan modifikasinya menggunakan *Attention Module* dengan Coordinate Attention (CA) yaitu CNN-CA-LSTM. Model yang dimodifikasi dengan Modul CA, dapat menghasilkan hasil prediksi atau *forecasting* yang lebih peka terhadap detail yang kecil [13].



Gambar 1. Metodologi Penelitian

Pada Gambar 1, menjelaskan tentang penelitian yang dilakukan saat ini dengan tahapan pengumpulan data curah hujan dari tanggal 1 Januari 1960 sampai 21 Mei 2026 dan peneliti mengambil data dari *pustaka open meteo* [3], komposisi dari data presipitasi curah hujan ini mengalami fluktuasi data yang tajam dikarenakan pola musiman pada cuaca Bojonegoro. Data akan melewati tahapan *data preprocessing* dengan menormalisasi data menggunakan Min-Max Scaler. Normalisasi data ini diperlukan untuk mengurangi dampak dari vanishing gradient pada model kecerdasan buatan yang digunakan. Penelitian ini mengeksplorasi metode Harmonic Dataset Distillation (HDD) yang berguna untuk mensintesis *dataset* menjadi lebih kecil dan ringkas untuk mencapai kinerja yang tidak jauh berbeda dengan *dataset* yang asli.

Penelitian ini dilakukan dengan membandingkan model dasar, yaitu menggunakan model CNN-LSTM dan CNN-CA-LSTM pada dataset asli, dan model usulan, yaitu model CNN-LSTM dan CNN-CA-LSTM, dengan dataset sintesis dari Harmonic Dataset Distillation (HDD), dengan eksplorasi model backbone, yaitu CNN-LSTM dan CNN-CA-LSTM. Penelitian ini menggunakan parameter evaluasi, yaitu RMSE, MAPE, R, R2, serta waktu pelatihan, yang akan digunakan untuk analisis komparatif pada model usulan dan model

dasar. Penelitian ini menggunakan metode evaluasi *Single Hold-Out* dengan membagi data menjadi 80 persen untuk data pelatihan dan 20 persen untuk data uji, serta matriks evaluasi RMSE, MAPE, R, dan R^2 dihitung pada data uji, serta pemilihan model untuk implementasi sederhana pada website berbasis Streamlit.

a. Pengumpulan Data

Pengumpulan data dilakukan dengan API dari Open Meteo dengan mengambil satu parameter yaitu data presipitasi curah hujan di Bojonegoro, dengan rentang waktu dari tanggal 1 Januari 1960 sampai 21 Mei 2026 dengan mengagregasi pada data setiap bulan, menghasilkan data sebanyak 797 data bulanan presipitasi curah hujan. Pustaka Open Meteo terhubung pada data ERA5 [3]. Pengambilan data presipitasi curah hujan sengaja dilakukan pada setiap bulan untuk menyaring fluktuasi data yang sangat tinggi. Hal ini dapat meningkatkan stabilitas pada pelatihan model dasar dan model usulan, serta mempersiapkan data pada Metode *Harmonic Dataset Distillation* (HDD).

Tabel 1. Statistik Deskriptif pada Dataset Curah Hujan Bojonegoro

Statistik	Nilai/Satuan
Jumlah Data	797 Sampel
Rata-Rata Data	175.55 mm
Simpangan Baku	147.49 mm
Nilai Minimum	0 mm
Kuartil 1 (Q1)	26.90 mm
Kuartil 2 (Q2)	155.70 mm
Kuartil 3 (Q3)	302.50 mm
Nilai Maksimum	579.90 mm

Pada Tabel 1, dapat diketahui untuk jumlah data yang digunakan ada 797 bulan, yang diambil mulai pada tahun 1 Januari 1960 sampai pada 21 Mei 2026. Dapat diketahui, pada Nilai minimum curah hujan per bulan, yaitu terdapat pada 0 mm, dan nilai maksimum untuk curah hujan bulanan adalah 579.90 mm yang merupakan curah hujan yang cukup tinggi dalam satu bulan, dan dapat diketahui terdapat fluktuasi yang tinggi pada data presipitasi curah hujan yang terjadi di Bojonegoro. Serta perlu diketahui, pada bulan Mei 2026 adalah data agregasi parsial pada Tanggal 1 Mei hingga 21 Mei 2026.

b. Data Pra-Pemrosesan

Tahap ini akan mempersiapkan data untuk dapat diolah dengan lebih baik pada model kecerdasan buatan. Pada penelitian ini, akan menggunakan analisis *Time series* pada data presipitasi curah hujan di Bojonegoro. Dapat diketahui berdasarkan Tabel 1, terdapat fluktuasi yang tinggi pada data presipitasi curah hujan yang terjadi di Bojonegoro, hal ini dapat meningkatkan potensi *Vanishing Gradient* saat pelatihan model yang dilakukan.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

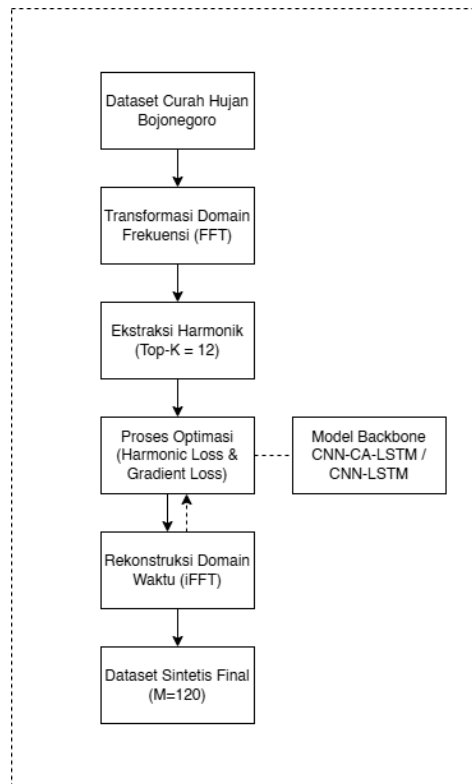
Pra-Pemrosesan data yang dilakukan pada penelitian ini, yaitu menggunakan metode Normalisasi data dengan *Min-Max Scaler* seperti pada persamaan 1, untuk mengurangi dampak *Vanishing Gradient* saat pelatihan model, serta *dataset* akan dibagi menjadi data pelatihan sebanyak 80 persen dan 20 persen adalah data uji untuk metode dasar dan untuk metode *Harmonic Dataset Distillation* akan menggunakan 80 persen data pelatihan untuk menghindari kebocoran data, serta penelitian ini menggunakan *time step* sebesar 12. Data untuk pelatihan bersifat kronologis dengan data pelatihan yaitu data riwayat curah hujan lama dan data uji merupakan data curah hujan yang terbaru.

c. Metode *Harmonic Dataset Distillation*

Pada tahap ini, Peneliti akan menggunakan *Harmonic Dataset Distillation* (HDD). Metode ini berfungsi untuk menyintesis atau distilasi dari *dataset* yang asli dan secara khusus untuk mendistilasi pada data *Time series*. Metode *Harmonic Dataset Distillation* berfungsi untuk membuat *dataset* sintesis dengan



meniru pola *dataset* asli, dengan tujuan *dataset* sintetis ini memiliki karakteristik yang hampir mirip dengan *dataset* asli namun dengan ukuran yang lebih kecil. Metode *Harmonic Dataset Distillation* ini dirancang secara khusus untuk mendistilasi tipe data *Time series*, yang berfungsi dengan mensintesis domain data waktu asli menuju domain frekuensi yang memiliki karakteristik yang hampir sama dengan *dataset* asli.



Gambar 2. Alur *Harmonic Dataset Distillation* (HDD)

Pada Gambar 2, Backbone = CNN-CA-LSTM dan CNN-LSTM, Panjang M=120, Top-K=12, Batch Size = 16, Iterasi = 200 dapat diketahui untuk proses yang terjadi pada *Harmonic Dataset Distillation*. Parameter yang digunakan yaitu M= 120 yang merepresentasikan *dataset* sintetis yang dihasilkan metode ini, Top-K=12 ditentukan untuk proses distilasi pada *dataset* asli yaitu untuk menangkap pola musiman dan mereduksi sinyal *noise* yang kurang esensial, *Batch Size* 16 dipilih secara khusus pada proses distilasi menggunakan HDD, dengan ukuran *batch size* ini dapat memberikan stabilitas estimasi gradien serta memberikan regularisasi stokastik yang baik saat proses distilasi dengan metode HDD. Proses distilasi diawali dengan mentransformasi pada data observasi presipitasi curah hujan di Bojonegoro dengan rumus *Fast Fourier Transform* (FFT).

$$F_x[k] = \sum_{n=0}^{M-1} x_n e^{-i2\pi kn/M} \quad (2)$$

Tahapan FFT menggunakan rumus pada Persamaan 2 ini, berfungsi untuk mendekomposisi pada data presipitasi curah hujan menjadi sekumpulan fungsi basis sinusoidal. Setelah proses transformasi Maka selanjutnya adalah proses ekstraksi harmonik pada data tersebut, dan pada penelitian ini menggunakan parameter Top K-12 dan tetapan M=120. Selanjutnya adalah tahap optimasi iteratif dengan model *backbone* oleh arsitektur CNN-CA-LSTM dan CNN-LSTM, dengan masing-masing percobaan sekali. Pada tahap ini model tersebut tidak bertujuan untuk memperbarui bobot jaringannya sendiri tetapi berfungsi sebagai evaluator untuk menghitung *Gradient Matching Loss*, algoritma ini mengukur pada jarak lintasan gradien model saat dilatih dengan data asli dengan lintasan gradien saat evaluasi pada data sintetis. Metode ini menghasilkan nilai *Gradient Matching Loss* (GML) dan *Harmonic Loss* (HL) sebagai nilai evaluasi *loss* dengan kriteria konvergensi adalah dari batas iterasi yaitu 200 dan kestabilan pada hasil total *Gradient*

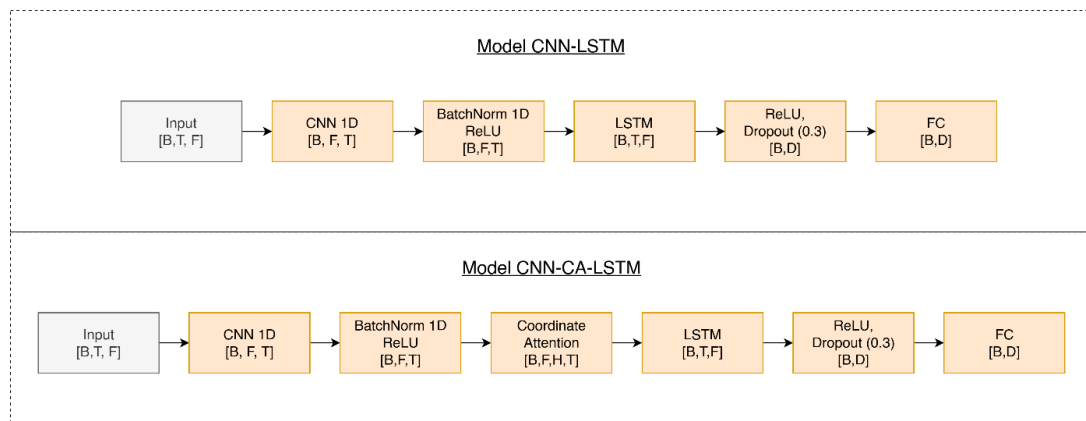
Matching Loss (GML) dan *Harmonic Loss* (HL). Melalui mekanisme *backpropagation*, sinyal kesalahan dari nilai *loss* tersebut digunakan untuk memperbaharui parameter pada *dataset* sintesis dengan bersamaan pada tahap rekonstruksi domain waktu, siklus ini akan dilakukan terus menerus hingga menghasilkan evaluasi *loss* yang konvergen.

$$X_H = iFFT(\tilde{F}_X), \quad S_H = iFFT(\tilde{F}_S) \quad (3)$$

Setelah hasil evaluasi pada data sintesis mencapai titik konvergen, maka akan dilakukan tahap Rekonstruksi Domain Waktu dengan rumus *inverse Fast Fourier Transform* (iFFT) dengan menggunakan tetapan $M=120$, menghasilkan *dataset* sintesis final yang berukuran lebih ringkas dengan hasil data *time step* sintesis namun memiliki karakteristik hampir sama dengan *dataset* asli.

d. Perancangan Model

Penelitian ini menguji coba pada dua jenis metode, yaitu metode dasar menggunakan Model CNN-LSTM dan CNN-CA-LSTM dengan *dataset* asli, serta metode usulan menggunakan kedua model yang sama namun dengan *dataset* sintesis hasil dari *Harmonic Dataset Distillation*. Peneliti menggunakan Framework PyTorch untuk membuat model ini. Berikut ini adalah penjelasan dari Model tersebut:



Gambar 3. Model CNN-LSTM dan CNN-CA-LSTM

Penelitian ini menggunakan *Time Step* sebesar 12, serta *Batch size* untuk pelatihan sebesar 8. Untuk rincian pada Gambar 3, dengan B adalah Jumlah sampel baris sekuens, T adalah *Time Steps*, dan F adalah Jumlah Fitur. Proses konvolusi dilakukan pada data 1 Dimensi yaitu pada data *time series* yang telah ditukar aksis dimensi yaitu pada dimensi T dan dimensi F. Pada Model CNN-CA-LSTM terdapat modul CA dengan *input* dimensi berjumlah 4 Dimensi, data dari BatchNorm 1D dan ReLU akan dilakukan Ekspansi Dimensi, untuk menyesuaikan ukurannya menjadi 4 Dimensi, data kemudian diolah dengan operasi *pooling* asimetris untuk mengekstraksi dari pola spasial berdasarkan hubungan antar kedekatan pada fitur pada data *time series* yang kemudian dilakukan penyatuan dan reduksi dimensi menjadi 3 Dimensi untuk olah data secara temporal dengan LSTM.

Metode dasar ini dilatih menggunakan *dataset* asli, serta memodelkan data menggunakan model CNN-LSTM dan CNN-CA-LSTM. Model CNN-LSTM adalah model gabungan dari model CNN dan LSTM. Model tipe ini juga digunakan pada analisis data *time series* [14], [15], [16], dengan tambahan modul CNN yang berfungsi untuk menambahkan kemampuan model dalam menangkap informasi spasial, yang digabungkan dengan modul LSTM untuk menangkap pola data temporal. Model CNN-CA-LSTM adalah hasil penggabungan antara model CNN-LSTM dengan modul Attention, yaitu Coordinate Attention (CA). Modul Coordinate Attention dapat meningkatkan kemampuan model, seperti meningkatkan representasi fitur yang ada, serta dapat memiliki kemampuan lebih dalam, seperti lebih mampu mengenali fitur-fitur spasial dan temporal [12].

Metode usulan dilatih menggunakan *dataset* sintesis dari *Harmonic Dataset Distillation*. Hasil *dataset* sintesis dari metode *Harmonic Dataset Distillation* meniru pola yang ada di *dataset* asli. *Dataset* hasil distilasi ini kemudian dilatih menggunakan modul CNN-LSTM dan CNN-CA-LSTM untuk

mengetahui perbedaan komparatif dari hasil distilasi dengan *dataset* asli. Metode usulan ini akan menggunakan model pemandu atau model *backbone* yaitu penelitian ini mengeksplorasi menggunakan CNN-LSTM dan CNN-CA-LSTM.

e. Parameter Evaluasi

Parameter evaluasi kuantitatif ini digunakan untuk mengevaluasi model hasil pelatihan pada analisis *Time series* ini. Model yang dilatih kemudian dievaluasi menggunakan parameter RMSE, MAPE, R, R^2 , Waktu Komputasi, serta *Mean Square Error* (MSE) pada Persamaan 8 sebagai fungsi *loss* untuk pelatihan model kecerdasan buatan dan memakai Validasi Hold-Out yang membagi data 80 persen untuk data pelatihan dan 20 persen untuk data uji pada metode dasar. Pemilihan matriks evaluasi ini berdasarkan dari fungsional matriks tersebut dalam mengevaluasi data *forecasting* curah hujan, seperti RMSE pada persamaan 4, untuk mengetahui bobot penalti saat evaluasi model pada fluktuasi ekstrem data curah hujan, R pada persamaan 6 untuk mengukur arah dan hubungan linear antar data aktual dan data hasil *forecasting*, R^2 pada persamaan 7 untuk mengukur dari variabilitas data curah hujan dari data asli dengan data hasil *forecasting*, dan MAPE pada Persamaan 5 yang memberikan gambaran umum pada hasil *forecasting* yang dilakukan.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} \quad (6)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (5)$$

$$R^2 = 1 - \frac{RSS}{TSS} \quad (7)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

Penelitian ini menggunakan data curah hujan yang mengalami banyak fluktuasi data dan mengalami inflasi data nol, matriks evaluasi MAPE akan menghasilkan persentase *error* yang tajam dan tinggi, MAPE akan tetap ditampilkan sebagai informasi tambahan serta tidak menunjukkan bahwa Model tersebut buruk, tetapi mengindikasikan bahwa ada inflasi data nol dan fluktuasi data yang tinggi di *dataset* curah hujan, dan matriks evaluasi utama adalah RMSE, R, R^2 , dan waktu komputasi untuk Metode Dasar dan Metode Usulan.

f. Implementasi Streamlit

Streamlit adalah pustaka yang banyak digunakan di lingkungan pemrograman bahasa Python untuk menerapkan dari model kecerdasan buatan dan ditampilkan dalam bentuk website. Streamlit memudahkan dari pengguna awam untuk dapat mengakses, mengenali, serta berinteraksi dengan Model kecerdasan buatan [17]. Pustaka Streamlit, bekerja di lingkungan pemrograman Python serta dirancang untuk dapat berinteraksi dengan pustaka-pustaka yang sering digunakan untuk kecerdasan buatan seperti Scikit-Learn, Numpy, dan PyTorch. Pustaka ini dapat meningkatkan dari kecepatan untuk pemrograman aplikasi dengan Graphic User Interface (GUI), karena fungsi-fungsi untuk membantu membuat *User Interface* sudah tersedia pada pustaka ini.

3. Hasil dan Pembahasan

a. Deskripsi Data

Data yang digunakan pada penelitian ini adalah data presipitasi curah hujan di Bojonegoro mulai tahun 1960 sampai dengan Mei 2026. Secara detail, data tersebut diambil setiap hari mulai tanggal 1960-01-01 sampai 2026-05-21, kemudian diagregasi per bulan untuk mengatasi gradien yang menghilang dikarenakan permasalahan banyaknya hari yang tanpa hujan. Data ini didapatkan dari Pustaka API Open-Meteo yang terhubung pada pustaka ERA5. Data yang diambil yaitu satu parameter yaitu parameter

presipitasi curah hujan dan data ini dipakai untuk proses pelatihan model kecerdasan buatan pada penelitian saat ini. Berikut ini Tabel yang menjelaskan cuplikan data tersebut:

Tabel 2. Data Presipitasi Curah Hujan di Bojonegoro

Tanggal	Presipitasi (mm)
1960-01-01	354.9
1960-02-01	327.4
1960-03-01	232.7
...	
2026-03-01	234.9
2026-04-01	272.7
2026-05-01	56.8

Berdasarkan pada Tabel 2. Dapat diketahui untuk data yang digunakan yaitu data agregasi per bulan dari data per hari, untuk data pada tanggal 2026-05-01 ini merupakan hasil agregasi mulai tanggal 2026-05-01 sampai 2026-05-21 yaitu hari saat data ini diambil. Data ini berisi pada data agregasi per bulan, dan untuk mengatasi banyaknya fluktuasi pada parameter presipitasi curah hujan, dilakukan agregasi data per bulan untuk merepresentasikan data curah hujan di Bojonegoro.

b. Analisis Performa Pelatihan Data

Tahap ini akan menjelaskan untuk hasil eksplorasi yang dilakukan di penelitian ini. Model kecerdasan buatan dan Metodologi yang dirancang digunakan untuk memudahkan mengenali dari data *Time series*. Penelitian ini menggunakan fungsi *Loss* MSE dari pustaka PyTorch. Penelitian ini menggunakan parameter yang tetap untuk setiap pelatihan pada Metode Dasar dan Metode Usulan. Berikut ini adalah detail pada *Hyperparameter* yang digunakan:

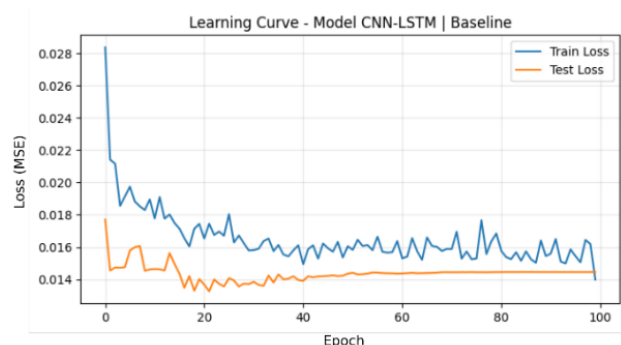
Tabel 3. Pengaturan *Hyperparameter*

<i>Hyperparameter</i>	Tetapan
<i>Optimizer</i>	Adam
<i>Weight Decay (Optimizer)</i>	0.0001
<i>Learning Rate</i>	0.001
<i>Batch Size</i>	8
<i>Time Step</i>	12
<i>ReduceLROnPlateau (Patience)</i>	5
<i>ReduceLROnPlateau (Factor)</i>	0.5
<i>Dropout</i>	0.3
<i>Random Seed</i>	42

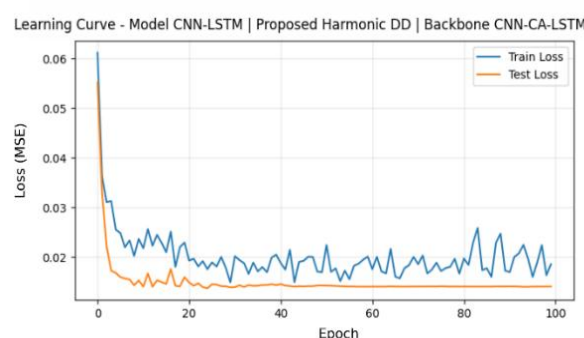
Pada Tabel 3, dijelaskan tentang *hyperparameter* yang digunakan selama pelatihan pada penelitian ini. Penelitian ini menggunakan *optimizer* Adam dan *learning rate* awal yaitu 0.001 dengan *scheduler* menggunakan *ReduceLROnPlateau* dengan tetapan *Patience* 5 dan Faktor 0.5 untuk membuat hasil pada *testing loss* menghasilkan evaluasi yang lebih konvergen dikarenakan tetapan parameter learning rate yang lebih dinamis. Penelitian ini menggunakan konfigurasi *weight decay* pada *optimizer* Adam sebesar 0.0001 dan *Dropout* sebesar 0.3 pada kedua Model untuk mencegah *overfitting*. Penelitian ini juga menggunakan tetapan *Random Seed* 42 pada pelatihan data *Time series* dan *Batch Size* sebesar 8 untuk pelatihan model pada Metode Dasar dan Metode Usulan, untuk Metode HDD menggunakan tetapan *Batch Size* sebesar 16 berbeda dengan tahapan Pelatihan yang diketahui dari Gambar 2, serta penerapan ini diterapkan pada pustaka Numpy, dan PyTorch. Penetapan *Random Seed* pada Pustaka yang digunakan untuk pelatihan model kecerdasan buatan dapat membantu untuk menghasilkan hasil evaluasi yang lebih stabil, tanpa perubahan yang signifikan pada hasil di parameter evaluasi.

c. Analisis Performa pada Fungsi Loss

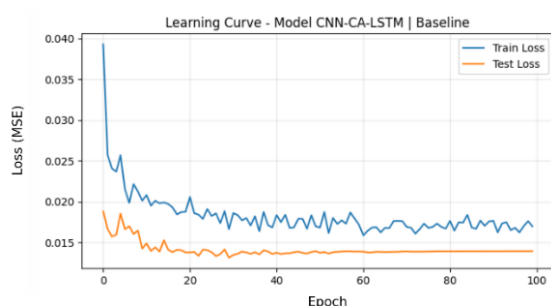
Pada tahap ini menjelaskan pada hasil fungsi *loss* pada pelatihan untuk analisis *Time series*. Berikut ini adalah gambaran dari hasil fungsi *loss* pada metode tersebut:



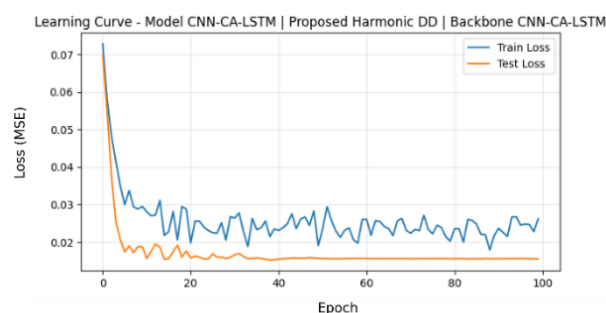
Gambar 4. Kurva Pembelajaran Model CNN-LSTM Pada Metode Dasar



Gambar 5. Kurva Pembelajaran Model CNN-LSTM dengan HDD dan Backbone CNN-CA-LSTM



Gambar 6. Kurva Pembelajaran Model CNN-CA-LSTM Pada Metode Dasar



Gambar 7. Kurva Pembelajaran Model CNN-CA-LSTM dengan HDD dan Backbone CNN-CA-LSTM

Berikut adalah analisis performa Model yang digunakan pada *forecasting* presipitasi curah hujan di Bojonegoro, pada metode dasar dan metode usulan. Dapat diketahui bahwa Gambar 4 dan Gambar 6 merupakan hasil kurva pembelajaran model CNN-LSTM pada Metode Dasar dan Metode Usulan dengan backbone CNN-CA-LSTM, yang terdapat pada Gambar 5 dan Gambar 7. Dapat diketahui, pada kelompok Metode dasar, pada Model CNN-LSTM dan CNN-CA-LSTM di metode dasar ini mengalami nilai konvergen pada Epoch ke 60, terdapat indikasi kenaikan *test loss* pada CNN-LSTM *epoch* 40-60 dan belum ditemukan indikasi *overfitting* yang signifikan.

Pada Gambar 5 dan Gambar 7, pada model CNN-LSTM dan CNN-CA-LSTM dengan backbone CNN-CA-LSTM diketahui terdapat fluktuasi hasil train loss pada epoch 80–100, tetapi model mampu mempertahankan hasil *test loss* dengan baik. Pada Gambar 6, Model CNN-LSTM dengan metode dasar juga mengalami sedikit kenaikan *test loss* pada *epoch* 40-80 sedangkan pada Metode Usulan tidak mengalami demikian. Metode HDD dengan *Backbone* CNN-CA-LSTM dapat membantu mensintesis data menjadi lebih mirip dengan *dataset* asli dengan representasi fitur yang lebih baik dikarenakan Modul CNN-LSTM dengan modifikasi CA dapat membantu menangkap informasi yang penting terhadap data [11], [12].

d. Analisis Komparatif

Analisis komparatif pada penelitian ini, yaitu mengeksplorasi pada metode dasar dan metode usulan. Metode dasar meliputi dari Model CNN-LSTM dan Model CNN-CA-LSTM, dan metode usulan dengan HDD. Analisis komparatif menggunakan *dataset* presipitasi curah hujan di Bojonegoro, kemudian data tersebut dilakukan pra-pemrosesan data dengan metode *Min-Max Scaler*, kemudian *dataset* dibagi menjadi 80 persen data pelatihan dan 20 persen data uji dan data yang telah di normalisasi ini akan menjadi *dataset* asli untuk pelatihan model. Metode usulan dengan *Harmonic Dataset Distillation* menggunakan 80 persen data pelatihan untuk di distilasi, hal ini dikarenakan untuk mencegah dari kebocoran data saat pelatihan. Berikut ini adalah hasil eksplorasi pada penelitian ini:

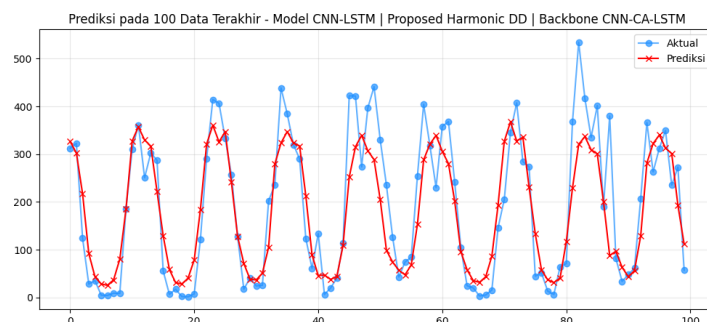
Tabel 4. Analisis Komparatif pada Metode Usulan dan Metode Dasar

Metode	Model <i>Backbone</i>	RMSE	MAPE	R	R ²	Waktu Komputasi (detik)
CNN-LSTM	-	69.71	267.60	0.8868	0.7742	38.80
CNN-CA-LSTM	-	68.38	253.97	0.8935	0.7827	59.49
CNN-LSTM + HDD	CNN-LSTM	72.57	319.98	0.8736	0.7553	10.30
CNN-CA-LSTM + HDD	CNN-LSTM	74.68	312.89	0.8821	0.7409	16.22
CNN-LSTM + HDD	CNN-CA-LSTM	68.92	305.61	0.8868	0.7793	11.05
CNN-CA-LSTM + HDD	CNN-CA-LSTM	72.24	311.64	0.8862	0.7575	15.67

Pada Tabel 4, diketahui untuk metode usulan dengan model CNN-CA-LSTM pada Metode dasar memiliki tingkat evaluasi yang paling baik di antara model lainnya dengan RMSE 68.38. Dapat diketahui, dari matriks evaluasi pada metode CNN-LSTM dengan HDD menggunakan model *backbone* CNN-CA-LSTM menghasilkan matriks evaluasi yang baik yaitu RMSE: 68.92, dibandingkan dari metode CNN-LSTM dengan HDD menggunakan model *backbone* CNN-LSTM yaitu 72.57.

Dapat diketahui pada Model CNN-CA-LSTM pada metode dasar memiliki hasil evaluasi yang paling baik dan Metode usulan dengan Model CNN-LSTM dengan Model *Backbone* CNN-CA-LSTM memiliki hasil akurasi yang baik dikarenakan Model CNN-CA-LSTM sebagai model *Backbone* pada HDD memiliki pengaruh yang baik dalam membentuk data sintesis *time series* pada Model HDD, dan metode HDD ini dapat memangkas waktu komputasi pada tahap pelatihan menjadi lebih cepat. Model CNN-LSTM dengan modifikasi CA memiliki kemampuan yang lebih baik dan membantu model untuk fokus pada informasi yang lebih penting terhadap data [11], [12], serta berdasarkan hasil eksplorasi Model CNN-LSTM yang menjadi model *backbone* belum dapat mengekstraksi fitur dengan baik pada Metode HDD.

Diketahui dari analisis komparatif metode HDD dapat mengurangi waktu komputasi pelatihan hingga 73.5 %, seperti pada Model CNN-CA-LSTM pada Metode Dasar dan Metode Usulan CNN-CA-LSTM dengan *backbone* CNN-CA-LSTM. Namun, jika dibandingkan secara kuantitatif dengan penelitian terdahulu [7], [8], yang belum menerapkan HDD belum dapat dilakukan dengan baik, dikarenakan perbedaan karakteristik *dataset*, serta tidak tercantumnya variabel evaluasi waktu komputasi pada artikel tersebut. Model CNN-LSTM dengan *Backbone* CNN-CA-LSTM diketahui mendapatkan hasil evaluasi yang baik dengan waktu komputasi pada pelatihan yang lebih ringan dari Model yang belum menerapkan Metode HDD. Nilai MAPE diketahui sangat tinggi, seperti pada model CNN-LSTM dengan *Backbone* CNN-CA-LSTM mendapatkan hasil MAPE sebesar 305.61, hasil ini tergolong tinggi tetapi ini membuktikan bahwa data yang digunakan untuk pelatihan model mengalami fluktuasi yang tajam dan pada *dataset* terdapat inflasi nilai nol atau banyak nilai nol.

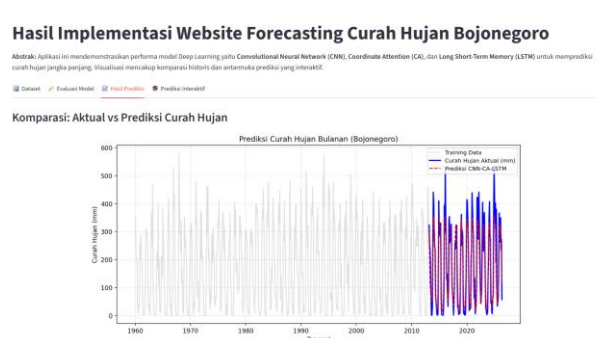


Gambar 8. Gambar Hasil Prediksi CNN-LSTM dengan Metode Usulan Model *Backbone* CNN-CA-LSTM

Dapat diketahui dari Gambar 8, hasil *forecasting* yang dibandingkan dengan data asli pada model CNN-LSTM dengan model *backbone* CNN-CA-LSTM. Model ini dapat mengenali pola pada 100 data uji terakhir dengan baik. Namun, beberapa kali model ini tidak mampu mengenali adanya fluktuasi data yang tinggi seperti pada bulan ke 80 sampai 90, model belumm mampu mengenali pola data dengan baik.

e. Hasil Implementasi Website

Antarmuka pada Streamlit berfungsi untuk interaksi pengguna aplikasi serta menampilkan hasil dari penelitian yang lebih interaktif. Pustaka Streamlit ini menyediakan pengalaman aplikasi *website* yang baik, intuitif, dan interaktif untuk dapat digunakan oleh pengguna.



Gambar 9. Halaman Dashboard Implementasi Website dengan Streamlite

Pada Gambar 9, dapat diketahui hasil implementasi website pada *forecasting* curah hujan di Bojonegoro dengan Model yang memiliki evaluasi terbaik yaitu CNN-CA-LSTM yang diterapkan pada Pustaka Streamlit. Antarmuka dirancang untuk berinteraksi secara langsung dan lebih mudah menggunakan aplikasi *website*, memastikan aksesibilitas bagi berbagai pengguna, termasuk pemangku kepentingan non-teknis. Fitur pada Tab pertama yang ditampilkan adalah Tab untuk Menampilkan *Dataset*, Fitur tab kedua adalah Evaluasi Model yang berisi Matriks Evaluasi seperti RMSE sampai R^2 dan menggunakan Min-Max Scaler, Tab ketiga berisi komparasi pada *Dataset* Asli dan Hasil *Forecasting*, dan Tab keempat mempunyai fitur untuk memperkirakan nilai presipitasi curah hujan pada bulan selanjutnya.

4. Kesimpulan dan Saran

Berdasarkan dari hasil eksplorasi yang dilakukan, dapat diketahui pada penerapan untuk metode usulan dan metode dasar terdapat perbedaan pada evaluasinya. Pada metode dasar yaitu menggunakan Model CNN-LSTM dan CNN-CA-LSTM dengan *dataset* asli mempunyai hasil evaluasi yang lebih baik seperti halnya pada hasil CNN-CA-LSTM dengan hasil RMSE paling kecil yaitu 68.38, namun memiliki waktu komputasi yang tertinggi. Metode Usulan menggunakan *Harmonic Dataset Distillation* diterapkan dengan mengeksplorasi menggunakan Model *Backbone* CNN-LSTM dan CNN-CA-LSTM. Dapat diketahui dengan menggunakan HDD hasil evaluasi cenderung terkoreksi mengalami penurunan kualitas, seperti pada model CNN-LSTM dengan model *backbone* CNN-CA-LSTM menghasilkan data sintesis memiliki tingkat evaluasi yang hampir sama dengan Metode dasar dengan Model CNN-LSTM, namun memiliki tingkat komputasi yang lebih ringan pada saat pelatihan terutama pada Metode Usulan dengan HDD. Model CNN-LSTM dengan HDD Model *Backbone* dengan CNN-CA-LSTM menghasilkan tingkat evaluasi yang lebih baik dengan RMSE 68.92 dari Model CNN-LSTM tanpa HDD dengan tingkat RMSE 69.71, hal ini diakibatkan oleh model CNN-LSTM memperoleh *dataset* sintesis dari Metode HDD dengan Model *Backbone* CNN-CA-LSTM untuk distilasi *dataset* asli, yang mengakibatkan hasil dari Metode *Harmonic Dataset Distillation* ini menghasilkan hasil *dataset* distilasi yang lebih baik dalam representasi data dan lebih membantu Metode HDD untuk fokus pada informasi yang lebih penting.

Dapat diketahui, pada penelitian ini memiliki keterbatasan, seperti menggunakan Metode Evaluasi *Single Hold-Out* dengan satu *Random Seed* yang belum memberikan gambaran variansi untuk analisis komparatif yang kuat, serta *zero inflated data* pada *dataset* presipitasi curah hujan. Tindakan yang dilakukan seperti melakukan agregasi data bulanan hanya mengurangi dari masalah inflasi data nol tetapi tidak menghilangkan dari dampak fluktuasi data dan dampak dari nilai nol tersebut terbukti dari hasil evaluasi MAPE yang tinggi. Penelitian selanjutnya dapat mengeksplorasi pada metode atau model yang dapat diterapkan pada permasalahan fluktuasi data yang tinggi, *zero inflated data* Seperti eksplorasi *Tweedie Loss* atau *Weighted Zero-inflated Sensitive Loss*, serta menerapkan *Repeated Hold-Out* dengan beberapa *seed* dan melaporkan pada rata-rata $\pm$ simpangan baku pada hasil evaluasi Model tersebut.

Referensi

- [1] H. Mulyanti, S. I. Purwaningrum, Y. L. A. D. Putra, A. A. Saputra, and B. Saputra, "Studi Adaptasi Perubahan Iklim pada Wilayah Pertanian Rawan Banjir di Kabupaten Bojonegoro," Bojonegoro, Feb. 2026. Accessed: Jun. 29, 2026. [Online]. Available: <https://repository.unigoro.ac.id/id/eprint/4966>
- [2] C. Sika Azkia, Y. Ansori, F. Darnis, and A. Rahman, "Pengaruh Feature Engineering Time Series Berbasis Rolling Statistics untuk Prediksi Suhu Menggunakan Random Forest dan Gradient Boosting," *Jurnal Informatika Universitas Pamulang*, vol. 11, no. 1, pp. 25–36, 2026, <https://doi.org/10.32493/jiup.v11i1.58137>.
- [3] A. Pico, M. A. Gonzalez-Sierra, and J. C. Martinez-Santos, "Machine Learning-Based Modeling of Reservoir Inflows from Precipitation Data," in *C3 2025 - IEEE Colombian Caribbean Conference*, Institute of Electrical and Electronics Engineers Inc., 2025. <https://doi.org/10.1109/C366505.2025.11339989>.
- [4] M. El Hafyani, K. El Himdi, and S. E. El Adlouni, "Improving monthly precipitation prediction accuracy using machine learning models: a multi-view stacking learning technique," *Frontiers in Water*, vol. 6, 2024, <https://doi.org/10.3389/frwa.2024.1378598>
- [5] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Machine Learning with Applications*, vol. 7, p. 100204, Mar. 2022, <https://doi.org/10.1016/j.mlwa.2021.100204>
- [6] E. G. A. Osman and F. A. Otaibi, "Integrating deep learning and econometrics for stock price prediction : A comprehensive comparison of LSTM , transformers , and traditional time series models," *Machine Learning with Applications*, vol. 22, no. September, p. 100730, 2025, <https://doi.org/10.1016/j.mlwa.2025.100730>
- [7] K. Indra Jaya, A. Mulyanti Suhartini, N. Jariyatin, and D. Septiani, "Prediksi Indeks Harga Saham Gabungan Indonesia Menggunakan Model CNN-CA-LSTM dan Aplikasi Website Streamlit," *Jurnal Mahasiswa Teknik Informatika*, vol. 10, no. 1, Feb. 2026. <https://doi.org/10.36040/jati.v10i1.16901>
- [8] A. Wijayanto, A. Sugiharto, and R. Santoso, "Identifikasi Dini Curah Hujan Berpotensi Banjir Menggunakan Algoritma Long Short-Term Memory (LSTM) Dan Isolation Forest," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 3, pp. 637–646, Jul. 2024, <https://doi.org/10.25126/jtiik.938718>
- [9] M. Bilgili, E. Pinar, and T. Durhasan, "Global monthly sea surface temperature forecasting using the SARIMA, LSTM, and GRU models," *Earth Sci. Inform.*, vol. 18, no. 1, Jan. 2025, <https://doi.org/10.1007/s12145-024-01585-z>
- [10] S. Hong, S. Jang, W. Kweon, S. Kim, G. Lee, and H. Yu, "Harmonic Dataset Distillation for Time Series Forecasting," *AAAI Conference on Artificial Intelligence*, 2026, <https://doi.org/10.1016/j.measurement.2024.116114>
- [11] W. Du *et al.*, "An improved ResNet method for urban flooding water depth estimation from social media images," *Measurement (Lond.)*, vol. 242, Jan. 2025, <https://doi.org/10.1109/CVPR46437.2021.01350>
- [12] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 13708–13717, 2021, <https://doi.org/10.1038/s41598-024-70117-y>
- [13] M. Khalid, S. Deivasigamani, S. V., and S. Rajendran, "An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, <https://doi.org/10.1109/ACCESS.2023.3245146>
- [14] Y. Cai, R. Zhao, H. Wang, L. Chen, Y. Lian, and Y. Zhong, "CNN-LSTM Driving Style Classification Model Based on Driver Operation Time Series Data," *IEEE Access*, vol. 11, pp. 16203–16212, 2023, <https://doi.org/10.1016/j.energy.2023.128701>
- [15] S. Pan *et al.*, "Oil well production prediction based on CNN-LSTM model with self-attention mechanism," *Energy*, vol. 284, no. January, p. 128701, 2023, <https://doi.org/10.1038/s41598-025-94239-z>



- [16] B. Ladjal *et al.*, “Hybrid deep learning CNN-LSTM model for forecasting direct normal irradiance: a study on solar potential in Ghardaia, Algeria,” *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-94239-z.
- [17] K. Muchtar, S. Member, Y. Hafifah, and A. Febriana, “Edge AI-Based Detection for Defective Coffee Beans Using Deep Learning and Streamlit Framework,” *IEEE Access*, vol. 13, no. April, pp. 67977–67992, 2025, <https://doi.org/10.1109/ACCESS.2025.3561189>

