

RAINFALL PREDICTION USING THE SARIMAX AND LSTM METHODS IN SEMARANG CITY

PREDIKSI CURAH HUJAN MENGGUNAKAN METODE SARIMAX DAN LSTM DI KOTA SEMARANG

Rudi Setyo Prihatin¹, Eri Zuliarso²

^{1,2}Jurusan Magister Teknologi Informasi, Fakultas Teknologi Informasi dan Industri
Universitas Stikubank Jl. Tri Lomba Juang Semarang, Indonesia
*rudisetyo0027@mhs.unisbank.ac.id*¹, *eri299@edu.unisbank.ac.id*²

Abstract - The purpose of this study is to predict the decade rainfall in Semarang City using two main methods, namely Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) and Long Short Term Memory (LSTM). The methodology of this study begins with data pre-processing, which includes data deletion analysis using *dropna* and data normalization using *Min-Max Scaling* to reduce scale the to between 0 and 1. The dataset is then divided into 80% training data and 20% test data. The validity of the data (X_{test} , Y_{test}) using the best 56-epoch data validation (val_loss) is better than the validity of the training data ($loss$). On the other hand, SARIMAX uses the (2,1,2), (2,1,2,36) model, and its validation techniques include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). Specifically, the RMSE of the LSTM model is 19.6, and its RMSE of the SARIMAX is 31.05. The MAE of LSTM is 15.0, SARIMAX is 24.5, and the R^2 of LSTM is 0.814, and SARIMAX is 0.52. Lower RMSE and MAE values indicate lower prediction errors, but a higher R^2 value of 1 indicates that LSTM can explain 81% of the actual data variation, which is better than SARIMAX, which is only about 52%. The main finding of this study is that the LSTM model performs better when recommending rainfall datasets.

Keywords: LSTM, SARIMAX, Rainfall, MAE, RMSE

Abstrak - Tujuan dari penelitian ini adalah untuk memprediksi curah hujan dasarian di Kota Semarang menggunakan dua metode utama yaitu *Seasonal Autoregressive Integrated Moving Average with Exogenous Variables* (SARIMAX) dan *Long Short Term Memory* (LSTM). Metodologi penelitian ini dimulai dengan pra-pemrosesan data, yang meliputi analisis hapus data menggunakan *dropna* dan normalisasi data menggunakan *Min-Max Scaling* untuk mengurangi skala menjadi antara 0 dan 1. Kumpulan data kemudian dibagi menjadi 80% data latih (*train*) dan 20% data uji (*test*). Validitas data (X_{test} , Y_{test}) menggunakan validasi data 56 *epoch* terbaik (val_loss) lebih baik daripada validitas data latihan ($loss$). Di sisi lain, SARIMAX menggunakan model (2,1,2), (2,1,2,36), dan teknik validasinya meliputi *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), dan *Koefisien Determinasi* (R^2). Secara spesifik, RMSE Model LSTM adalah 19,6, dan RMSE SARIMAX-nya adalah 31,05. MAE LSTM adalah 15,0, SARIMAX adalah 24,5, dan R^2 LSTM adalah 0,814 serta SARIMAX adalah 0,52. Nilai RMSE dan MAE yang lebih rendah menunjukkan galat prediksi yang lebih rendah, tetapi nilai R^2 yang lebih tinggi, yaitu 1, menunjukkan LSTM dapat menjelaskan 81% variasi data aktual, yang lebih baik daripada SARIMAX yang hanya sekitar 52%. Temuan utama dari studi ini adalah bahwa model LSTM berkinerja lebih baik ketika merekomendasikan dasarian hujan curah..

Kata Kunci - LSTM, SARIMAX, Curah Hujan, MAE, RMSE

I. PENDAHULUAN

Di Indonesia, bencana hidrometeorologi dan perubahan iklim merupakan masalah serius [1]. Karakteristik iklim di Indonesia sangat dipengaruhi oleh sistem monsun, yang menciptakan dua musim utama yaitu musim hujan dan kemarau. Sebagai ibu kota Provinsi Jawa Tengah, Kota Semarang merupakan salah satu dari sedikit wilayah metropolitan yang mengalami tantangan hidrometeorologi. Letak geografisnya di pesisir utara Jawa menjadikan kota ini sangat menarik terkait dengan banjir, yang memengaruhi aktivitas sosial, ekonomi, dan transportasi masyarakat [2]. Untuk memprediksi curah hujan secara akurat, metode yang sudah ada seringkali menghasilkan keterbatasan. Menurut penelitian sebelumnya, model *Autoregressive Integrated Moving Average* (ARIMA) cukup baik dalam memprediksi masa depan, tetapi kurang baik dalam memodelkan hubungan non-linier yang kompleks[3].

Metode yang digunakan dalam makalah ini meliputi SARIMAX dan LSTM, yang dapat menghasilkan prediksi *non-linier*. Dalam penelitian [4] sebelumnya menggunakan model SARIMAX, curah hujan bulanan ditentukan sebesar 92,70 pada evaluasi MAE dalam sampel dan 86,22 pada sampel luar, sedangkan nilai MSE untuk sampel dalam adalah 15,56 dan sampel luar adalah 110,89. Akurasi prediksi kurva hujan menggunakan metode ARIMA dan LSTM ditunjukkan oleh nilai RMSE untuk LSTM dan SARIMA masing-masing sebesar 27,68 dan 28,36, sedangkan nilai MAE untuk LSTM dan SARIMA adalah 5,65 dan 5,79 dalam penelitian ini. Model LSTM lebih berhasil dalam memprediksi kurva hujan [5]. Penelitian ini adalah untuk mengembangkan prediksi curah hujan dasarian yang lebih akurat dan mumpuni menggunakan metode SARIMAX dan LSTM. Dengan mengetahui model mana yang terbaik, selanjutnya digunakan untuk membuat prediksi curah hujan dasarian. Dengan memprediksi hujan dasarian maka awal musim hujan dan awal musim kemarau dapat ditentukan [6].

II. SIGNIFIKANSI STUDI

A. Studi Literatur

Menggunakan metode SARIMAX untuk menentukan curah hujan dasarian masih cukup sulit. SARIMAX digunakan dalam penelitian [4] untuk menentukan curah hujan bulanan Tasikmalaya. Menggunakan model SARIMAX (0,0,1), prediksi terbaik dari hujan dibuat dengan evaluasi MAE sebesar 92,70 untuk sampel dan 86,22 untuk sampel luar, sedangkan nilai MSE untuk sampel dalam adalah 15,56 dan nilai MSE untuk sampel luar adalah 110,89. Dalam penelitian sebelumnya menggunakan LSTM untuk memprediksi curah hujan dasarian Stasiun Meteorologi Ngurah Rai (Badung Bali), RMSE evaluasi adalah 31,15 dan nilai R^2 adalah 0,53[7]. Setiap metode memiliki kelebihan dan kekurangan yang spesifik dalam konteks memprediksi curah hujan Kota Semarang. Metode SARIMAX sangat sesuai untuk memodelkan pola musiman yang regular menjadikannya efektif untuk prediksi jangka panjang. Namun, kelemahan utamanya adalah kurang akurat saat memprediksi curah hujan ekstrem[8]. Di sisi lain LSTM menunjukkan keunggulan dalam memodelkan data sekuensial misalnya curah hujan, mampu mengenali hubungan serta pola jangka panjang yang kompleks. Akan tetapi, performa optimal LSTM sangat bergantung pada ketersediaan data historis deret waktu yang panjang untuk proses pelatihannya[9].

Dengan adanya prediksi hujan dasarian, prediksi ini berfungsi sebagai alat untuk pengembangan keputusan strategis di Semarang. Bagi pemerintah daerah, khususnya BPBD, informasi ini berfungsi sebagai dasar untuk antisipasi perubahan musim dan meningkatkan efektivitas sistem peringatan bencana. Di sektor pertanian, prediksi ini memungkinkan petani untuk menyesuaikan pola tanam sesuai musim. Data prediksi hujan dasarian juga memberikan panduan teknis untuk pembangunan infrastruktur, khususnya di bidang mitigasi kekeringan dan bencana banjir.

B. Data Yang di Gunakan

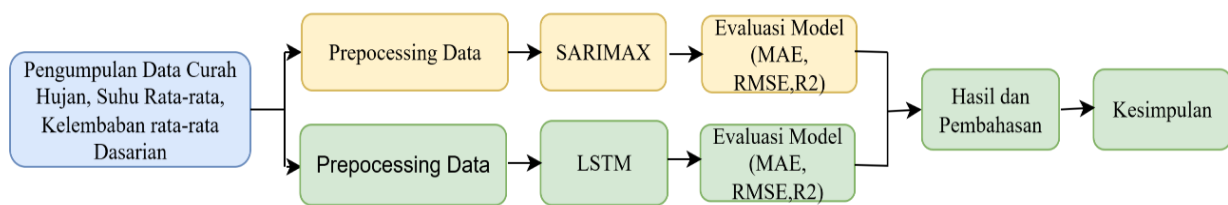
Langkah pertama dalam penelitian ini adalah mengumpulkan dan mempersiapkan data. Data yang digunakan merupakan data sekunder berjenis deret waktu (*time series*) yang di peroleh dari sumber yang kredibel Stasiun Klimatologi Jawa tengah, yang berada di bawah naungan Badan Meteorologi, klimatologi dan Geofisika (BMKG). Data mencakup rentang yang panjang yaitu 36 tahun dari Januari 1988 hingga Desember 2023. Periode yang panjang sangat relevan karena memungkinkan model untuk menangkap berbagai pola , baik pola musiman (msim hujan dan kemarau) yang meliputi data curah hujan dasarian, suhu rata-rata dasarian dan kelembaban udara rata-rata dasarian (10 harian), dengan jumlah data keseluruhan dalam penelitian ini 3885 baris data. Dataset ini di anggap cukup besar dan *representative* untuk melatih dan menguji model prediksi.

C. Lokasi Penelitian

Lokasi penelitian ini berada di Kota Semarang. Dengan karakteristik wilayah topografinya yang khas dimana di utara merupakan dataran rendah dan di selatan merupakan dataran tinggi[10].

D. Metode Penelitian

Metodologi Penelitian menggunakan model SARIMAX dan LSTM, seperti yang ditunjukkan pada gambar 1 :



Gambar 1. Tahapan Penelitian sarimak dan lstm

1. Preprocessing data

- a. *Cleaning Data*, Data perlu diperbaiki untuk menghasilkan keluaran data yang akurat. Menurut [11] *cleaning data* merupakan langkah yang penting sebelum menghasilkan output yang akurat. Jika data tidak akurat maka model dalam membuat prediksi curah hujan hasilnya kurang maksimal.
- b. *Smoothing Data*, Data Curah hujan adalah kumpulan data yang sangat bervariasi yang mencakup nilai yang tinggi dan terendah. *Smoothing data* digunakan untuk mengubah data menjadi bentuk yang lebih halus[12]. Data yang halus dianalisis menggunakan metode yang menghitung rata-rata tertentu dengan menganalisis data bergerak[13]. Formula untuk menghitung rata-rata bergerak[14]:

$$S_t = \frac{1}{3}X_{t-1} + X_t + X_{t+1} \quad (1)$$

S_t adalah nilai *Smoothing* yang baru , X_t adalah Nilai aktual waktu ke t, X_{t-1} adalah nilai aktual periode sebelumnya sedangkan X_{t+1} adalah nilai aktual sesudahnya.

Data curah hujan yang sudah di haluskan sangat baik digunakan untuk prediksi menggunakan SARIMAX dan LSTM.

- c. *Membagi data train dan test*, Untuk tahap selanjutnya data di bagi menjadi train 80 % dan data test 20 %. menggunakan fungsi *train-test-split*. Pembagian data memungkinkan validasi kinerja model pada set data validation yang terpisah. Dengan melakukan tahapan ini data telah dipersiapkan dengan baik sehingga dapat digunakan dalam training dan validation model.[15]

- d. *Normalisasi Data*, *Normalisasi data* menggunakan *minmaxscaller* dalam rentang -1 hingga 1 dengan persamaannya sebagai berikut [16]:

$$X_{norm_i} = \left(\frac{x_i - x_{min_i}}{x_{max_i} - x_{min_i}} \right) x2 - 1 \quad (2)$$

X_{norm} adalah variable yang dinormalisasi, x_{min} dan x_{max} masing – masing nilai variable minimum dan maksimum, dan i mewakili nilai untuk setiap titik data.

2. SARIMAX

Model SARIMAX adalah pengembangan dari model SARIMA yang melibatkan *variabel eksogen* yang dianggap memiliki pengaruh pada *variabel endogen*. Menurut Chen [17] persamaan SARIMAX secara umum dituliskan:

$$SARIMAX(p, d, q)(P, D, Q) = SARIMA(p, d, q)(P, D, Q) + \text{Variabel Eksogen} \quad (3)$$

Persamaan model SARIMA tanpa variabel eksogen :

$$\phi_p(L)\phi_p(L^S)(1-L)^d(1-L^S)^D Y_t = \theta_q(L) \ominus Q(L^S)\epsilon_t \quad (3)$$

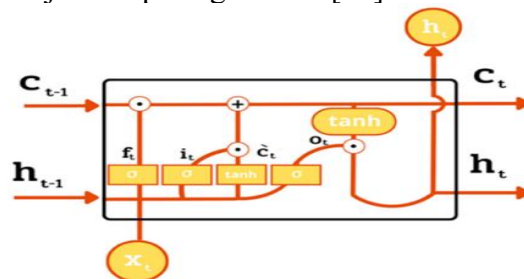
Persamaan SARIMA di tambah variabel eksogen (X)

$$\phi_p(L)\phi_p(L^S)(1-L)^d(1-L^S)^D Y_t = \theta_q(L) \ominus Q(L^S)\epsilon_t + \beta X_t \quad (4)$$

Y_t adalah data time series waktu t , L adalah *operator lag*, $\phi_p(L)$ adalah *polynomial AR non musiman* (orde p). $\phi_p(L^S)$ adalah *polynomial AR musiman* orde (P), $\ominus Q(L^S)$ adalah *polynomial MA musiman* orde (Q), d adalah *orde differencing non-musiman*, D adalah *orde differencing musiman*, S adalah *panjang siklus musiman* (misal sebulan ada 12 data), ϵ_t adalah *error / white noise*, X_t adalah *matriks variable eksogen* pada waktu t , β adalah koefisien regresi untuk *variable eksogen*.

3. Long Short - Term Memory (LSTM)

LSTM adalah jaringan saraf tiruan yang dirancang untuk menganalisis pola hubungan menggunakan data historis yang panjang. Struktur LSTM terdiri dari tiga lapisan neuron yang saling berhubungan erat dalam suatu struktur. Ada tiga jenis gerbang utama dalam unit LSTM yaitu input, output, lupa. Seperti yang di tunjukkan pada gambar 2[17].



Gambar 2. Struktur Metode LSTM (sumber Chen at all.,2022)

Model LSTM ditunjukkan pada gambar 2 dengan menyesuaikan gerbang input, output dan lupa. Dalam prediksi deret waktu *autokorelasi* deret waktu, hal ini sangat penting. *Autokorelasi* adalah dasar dari Model prediksi LSTM dan semakin besar *koefisien autokorelasi* (ACF) maka akan semakin baik[18]. Informasi yang pertama melalui blok memori adalah gerbang lupa dengan persamaan :

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

f_t adalah output dari gerbang lupa pada waktu t . Nilainya antara 0 dan 1 karena dihasilkan oleh fungsi aktivasi *sigmoid* (σ). Nilai 0 mempunyai berarti "lupakan semua informasi", sedangkan nilai 1 berarti "pertahankan semua informasi", σ adalah fungsi aktivasi *sigmoid*, yang didefinisikan sebagai berikut :

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (6)$$

Fungsi ini mengubah input menjadi rentang antara 0 dan 1, W_f adalah matriks bobot yang dipelajari selama training untuk gerbang lupa. Matriks ini menghubungkan input (x_t) dan *hidden state* sebelumnya (h_{t-1}) ke gerbang lupa, $[h_{t-1}, x_t]$ adalah gabungan dari *hidden state* sebelumnya (h_{t-1}) dan input pada waktu t (x_t), b_f adalah bias yang dipelajari untuk gerbang lupa, h_{t-1} adalah *hidden state* pada waktu sebelumnya $t - 1$, x_t adalah Input pada waktu t
Gerbang Masukan :

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (7)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (8)$$

i_t adalah output dari gerbang masukan pada waktu t . Nilainya antara 0 dan 1 karena dihasilkan oleh fungsi aktivasi sigmoid (σ). Nilai 0 mempunyai arti "abaikan informasi baru", sedangkan nilai 1 berarti "simpan semua informasi baru", σ adalah fungsi aktivasi *sigmoid*, W_i adalah matrik bobot yang dipelajari selama training untuk gerbang masukan. Matriks ini menghubungkan input (x_t) dan *hidden state* sebelumnya (h_{t-1}) ke gerbang masukan, $[h_{t-1}, x_t]$ adalah gabungan dari *hidden state* sebelumnya (h_{t-1}) dan input pada waktu t (x_t) adalah bias yang dipelajari untuk gerbang masukan, $\tilde{C}_t$ adalah kandidat nilai baru yang akan di tambahkan ke state sel. Nilai ini dihasilkan oleh fungsi aktivasi *tanh*, yang mengubah input menjadi rentang antara -1 dan 1, *tanh* adalah fungsi aktivasi *hiperbolik tangen*, yang didefinisikan sebagai berikut

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (9)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (10)$$

C_t adalah *Current state* (state sel pada awal waktu t), merupakan memory internal LSTM yang menyimpan informasi jangka panjang, f_t adalah Output dari gerbang lupa (*forget Gate*) pada waktu t nilainya Antara 0 dan 1, yang dihasilkan oleh fungsi *sigmoid* σ , C_{t-1} adalah state sel pada waktu sebelumnya $t - 1$ merupakan memori internal dari langkah sebelumnya i_t adalah output dari gerbang masukan pada waktu t . Nilainya antara 0 dan 1 karena dihasilkan oleh fungsi aktivasi *sigmoid* σ , $\tilde{C}_t$ adalah kandidat nilai baru yang akan ditambahkan ke state sel. Nilai ini dihasilkan oleh fungsi aktivasi *tanh*, yang mengubah input menjadi rentang -1 dan 1. Sedangkan gerbang keluaran terdiri dari dua langkah yaitu menghitung *output gate* (o_t) dan menghitung *hidden state* (h_t) :

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (12)$$

4. Evaluasi Model SARIMAX dan LSTM

Evaluasi model SARIMAX dan LSTM dilakukan menggunakan *Root Mean Square Error (RMSE)*, *Mean absolute error (MAE)* dan nilai R kuadrat untuk menilai seberapa baik model memprediksi. Menurut [19], RMSE adalah akar kuadrat dari rata-rata kuadrat kesalahan, metrik ini membantu mengukur tingkat kesalahan (*error*) antara nilai yang diprediksi dan nilai aktual yang diamati. Evaluasi dilakukan dengan membandingkan hasil, semakin rendah nilai RMSE maka semakin baik kinerja model. Menurut [20] MAE merupakan metrik evaluasi untuk mengukur rata – rata kesalahan absolut (perbedaan absolut) antara nilai yang prediksi dan nilai yang diamati sebenarnya. Semakin akurat model dalam prediksinya, semakin rendah nilai MAE. *Koefisien determinasi* (R^2)

Adalah metrik evaluasi yang di gunakan untuk mengevaluasi beberapa model regresi, seberapa baik model regresi menjelaskan variasi data aktual. Semakin dekat nilai R^2 ke 1 berarti semakin baik kinerja model. Perhitunganya adalah sebagai berikut[21]:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (13)$$

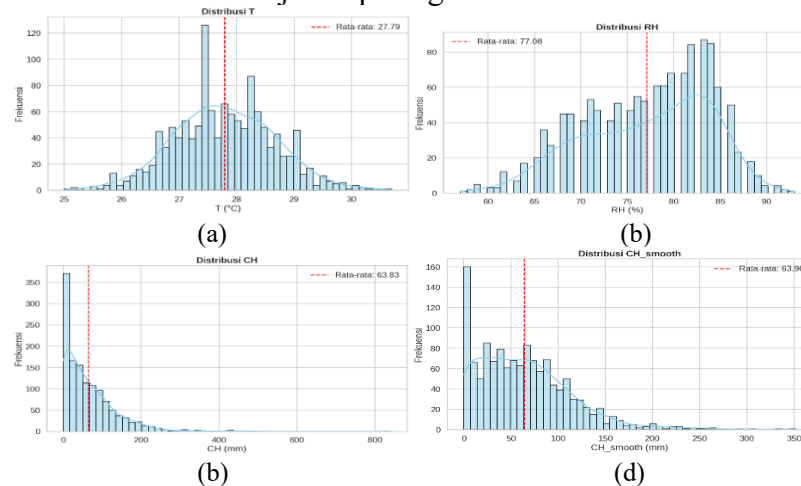
$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y(i) - \hat{y}_i)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2} \quad (15)$$

n adalah Jumlah titik yang diamati, y_i adalah Nilai aktual yang diamati pada titik ke i , $\hat{y}_i$ adalah Nilai yang prediksi pada titik i , $\sum$ = Notasi penjumlahan dari $i = 1$ hingga n , $|y_i - \hat{y}_i|$ adalah Nilai *Absolut* (selisih tanpa tanda negatif), $\bar{y}$ = rata-rata dari semua nilai aktual.

III. HASIL DAN PEMBAHASAN

Dengan menggunakan SARIMAX dan LSTM, analisis data curah hujan, suhu rata-rata, dan kelembaban rata-rata dasarian dari tahun 1988 sampai 2023 menghasilkan prediksi curah hujan dasarian. Analisis data-data tersebut ditunjukkan pada gambar 3 :



Gambar 3. Grafik Histogram Distribusi Data Suhu (a), Kelembaban (b), Curah Hujan (c), Curah Hujan Smooting (d)

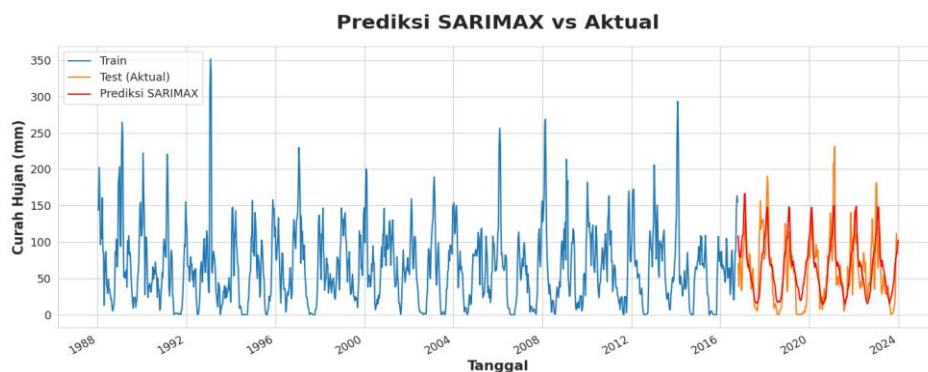
Empat histogram yang ditunjukkan pada Gambar 3 adalah data Suhu (T), Kelembaban (RH), Curah Hujan (CH), dan Curah Hujan *Smooth* (CH_smooth). Histogram ini menunjukkan beberapa nilai-nilai berulang yang muncul. Grafik suhu (a) menunjukkan simetri yang kurang lebih sama dengan distribusi normal. Berdasarkan data, suhu paling sering terlihat pada suhu antara 27 dan 28 derajat Celcius. Distribusi data lebih condong ke kiri pada kelembaban udara (b), dengan distribusi data sekitar 70% hingga 85% yang menunjukkan bahwa kondisi lembap sering terjadi.

Curah hujan aktual (c) sebarannya tidak konstan dengan kanan. Puncak grafik yang sangat tinggi dekat dengan nilai nol memiliki makna banyak hari tanpa hujan atau sedikit hujan, yang merupakan salah satu kejadian yang paling sering terjadi. Rentang data hujan sebenarnya yang berkisar antara 0 hingga 800 mm tergolong agak lebar. Rata-rata curah hujan sebenarnya adalah 63,8 mm. Curah hujan *Smooth* (d) pendistribusian yang sudah diperhalus condong ke kanan tetapi tidak ssebanyak

curah hujan aktual. Tujuan dari data curah yang telah mengalami proses smoothing adalah untuk meredam fluktuasi yang ekstrim. Hasil smoothing ini menunjukkan bahwa data curah hujan tinggi-tinggi telah berkurang terlihat puncak di dekat nilai 0 tidak sebanyak curah hujan aktual. Curah hujan rentang nilai datanya berkisar antara 0 hingga 350 mm, dengan rata-rata smoothing sebesar 63,9 mm. Hal ini lebih mudah untuk disimulasikan karena variasi data curah hujan lebih terasa.

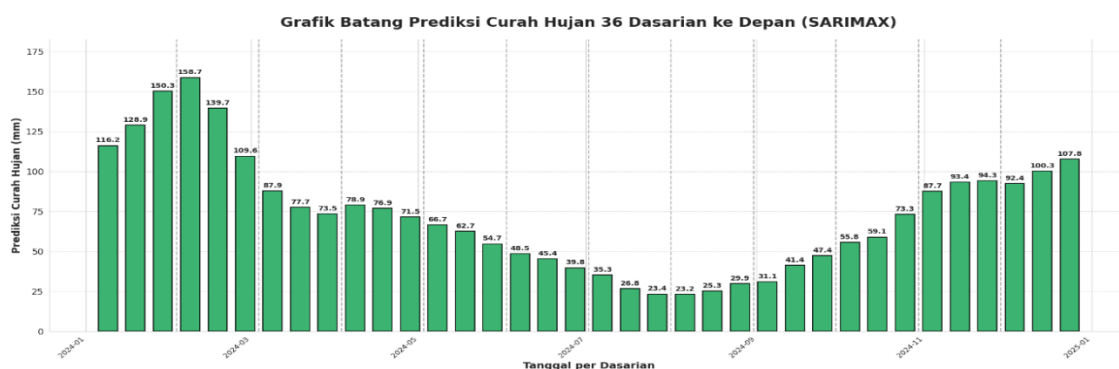
A. SARIMAX

SARIMAX adalah model statistik yang bekerja dengan baik dengan data deret waktu. Data suhu dan kelembaban digunakan sebagai eksternal digunakan untuk membantu meningkatkan akurasi, data pelatihan adalah data utama yang akan dievaluasi (*Endogen*). Dua jenis data yang akan digunakan membuat prediksi adalah data pelatihan dan data uji. Sekitar 80% data digunakan untuk melatih model agar dapat mempelajari pola dari data. Sekitar 20% data digunakan untuk pengujian guna membantu mengidentifikasi model yang baik untuk membuat prediksi. Model SARIMAX yang digunakan untuk memprediksi hujan dasarian adalah (2,1,2), (2,1,2,36).



Gambar 4. Grafik Train, Test dan Prediksi SARIMAX

Pada gambar 4 terlihat SARIMAX (merah) mampu menangkap pola musiman atau naik – turunya curah hujan dengan cukup baik. Ini menunjukkan bahwa komponen musiman bekerja dengan benar dalam mengidentifikasi kapan musim hujan dan kapan musim kemarau. Kesalahan SARIMAX dilihat pada gambar 4 saat prediksi puncak curah hujan tahun 2021-2022 model SARIMAX (oranye) memprediksi adanya puncak curah hujan pada periode tersebut, tetapi tingginya jauh dibawah aktualnya. Model SARIMAX dalam hal ini menunjukan *underfitting*. Model SARIMAX tidak amampu memprediksi curah hujan yang ekstrem. Prediksi curah hujan 36 dasarian kedepan yaitu tahun 2024 dapat dilihat pada gambar 5 :



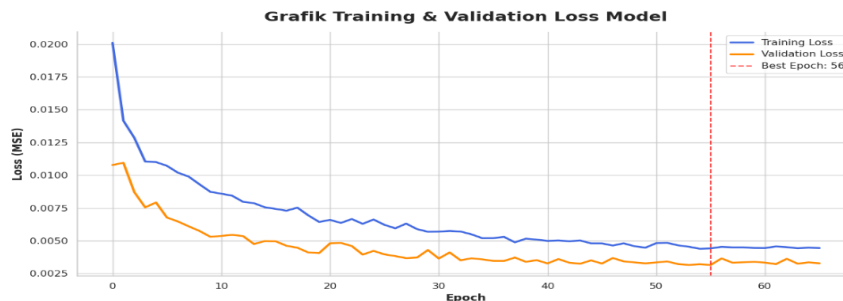
Gambar 5. Grafik Prediksi Curah Hujan 36 Dasarian

Pada Gambar 5 menunjukkan Prediksi curah hujan dasarian yang kurang dari 50 mm mulai bulan Juni dasarian 1 yaitu 48.5 mm hingga bulan September dasarian 3 yaitu 47.4. Curah hujan terendah

pada Juli dasarian 3 dan bulan Agustus dasarian 1 sebesar 23 mm. Sebaliknya curah hujan yang lebih dari 50 mm terjadi pada bulan Oktober dasarian 1 yaitu 55.8 mm hingga bulan Mei dasarian 3 sebesar 54.7. Curah hujan tertinggi terjadi pada dasarian 1 bulan Februari sebesar 158 mm.

B. LSTM

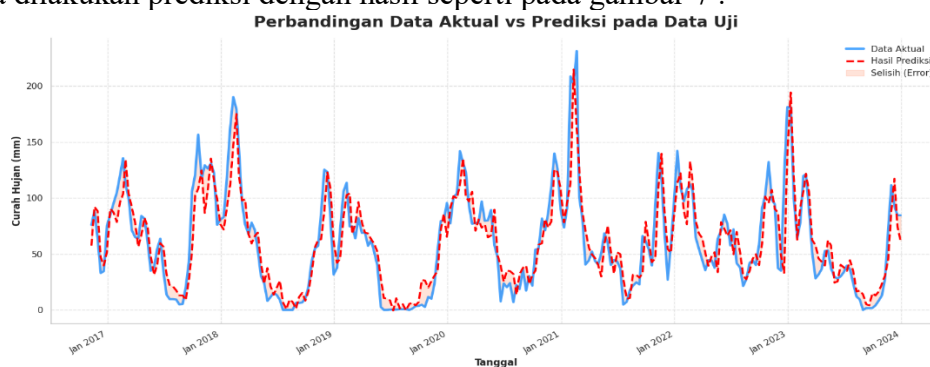
LSTM merupakan model *deep learning* yang sangat baik digunakan untuk membuat prediksi dengan data berurutan yang memiliki memori jangka panjang yang mampu mengenali pola prediksi jangka panjang. Seperti yang ditunjukkan gambar 6 adalah hasil permodelan LSTM penilaian kinerja model.



Gambar 6. Grafik Nilai *Loss* dan *Validasi Loss*

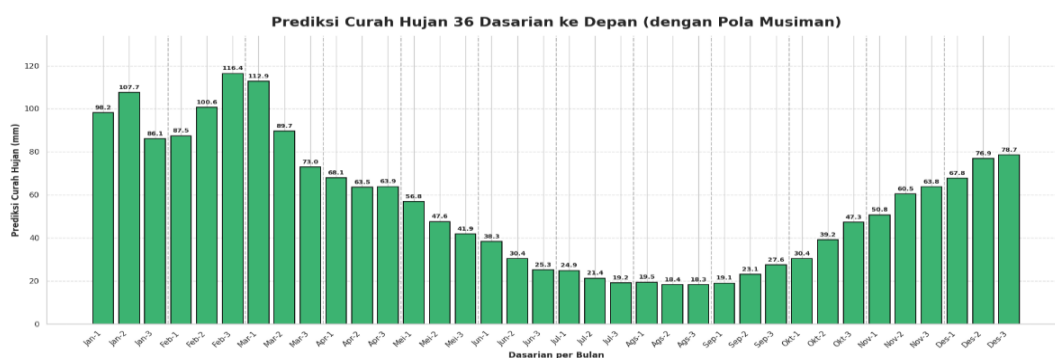
Pada pengujian model lstm untuk nilai *loss* dan *validasi loss* menggunakan *epoch* 100 dan *bach_zise* 16 dengan *data train* 80% dan *data test* 20%. Pada gambar 6 menunjukkan grafik nilai *loss* dan *validasi loss* sumbu x (*epoch*) merupakan siklus untuk pelatihan. Satu *epoch* berarti model telah memproses seluruh data latih sebanyak satu kali. Sedangkan sumbu Y (*loss*) mempresentasikan tingkat kesalahan model dimana semakin rendah nilainya berarti model semakin baik. Dalam grafik ini terdapat dua garis yaitu warna biru (*trainning loss*) yang menunjukkan tingkat kesalahan model pada data yang sedang di latih, sedangkan warna oranye (*validation loss*) menunjukkan tingkat kesalahan model pada data baru (data uji/validasi).

Berdasarkan grafik tersebut dapat disimpulkan bahwa model dapat belajar dengan baik hal ini terlihat dari kedua garis (biru dan oranye) mnunjukkan penurunan di awal kemudian melandai. Dalam hal ini model menunjukkan mampu mempelajari data dan kinerjanya meningkat pesat di awal kemudian menunjukkan stabil setelah menemukan pola yang relevan. Pola yang terbaik adalah pada *best epoch* 56. Jika dilihat dari kinerja model dapat disimpulkan bahwa model tidak terjadi *overfitting* (*model good fit*) dimana model tidak hanya menghafal data latih tetapi juga mampu menggeneralisasi pengetahuannya dengan baik ke data yang baru. Setelah model sudah terlihat setabil maka dilakukan prediksi dengan hasil seperti pada gambar 7 :



Gambar 7. Grafik Perbandingan Data Aktual dan Prediksi Curah Hujan

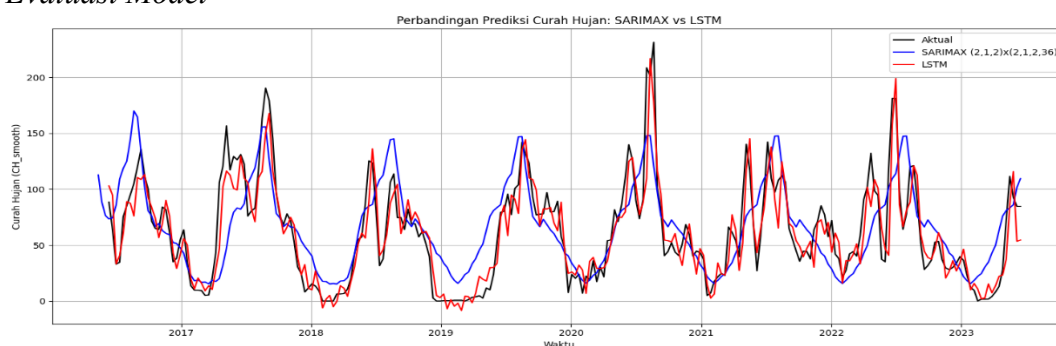
Gambar 7 menunjukkan model LSTM (merah) secara konsisten mengikuti pola dari data actual, hal ini menunjukkan model LSTM berhasil mempelajari siklus musiman. Model ini juga menunjukkan fleksibilitas dalam menangkap variasi dari tahun ke tahun. Model ini mampu mempelajari hubungan *non linier*, sebagai *neural network* mampu menangkap hubungan yang rumit secara implisit dari data. Walaupun model ini sangat baik untuk prediksi tetapi ada kekurangannya, hal ini dapat ditunjukkan pada prediksi tahun 2021 model LSTM (merah) berhasil memprediksi adanya puncak periode yang sama namun tingginya masih dibawah nilai aktualnya. Model kesulitan untuk mengukur kejadian curah hujan yang ekstrem. Hal ini dapat terjadi karena kejadian hujan ekstrem mungkin jarang terjadi dalam data latih. Model juga sensitif terhadap *outliers* model ini dilatih pada data historis. Jadi jika di masa depan terjadi *anomali* cuaca yang belum pernah ada sebelumnya, model mungkin gagal memprediksi secara akurat. Adanya resiko *overfitting* bahwa model menjadi terlalu hafal dengan pola data latih dan kehilangan kemampuan untuk menggeneralisasi pada data baru. Hasil prediksi model LSTM untuk 36 dasarian ke depan dapat dilihat dari gambar 8 :



Gambar 8 Grafik Prediksi Curah Hujan dasarian menggunakan LSTM

Grafik pada gambar 8 prediksi curah hujan di bawah 50 mm mulai bulan Mei dasarian 2 yaitu 47.6 hingga bulan Oktober dasarian 3 sebesar 47.3 mm. Dasarian ketiga bulan Agustus Curah hujan terendah sebesar 18.3 mm. Curah hujan lebih dari 50 mm bulan November dasarian 1 tercatat curah hujan 50.8 mm hingga bulan Mei dasarian 1 sebesar 56.8 mm. Curah hujan tertinggi terjadi pada bulan Februari dasarian 3 sebesar 116.4 mm.

C. Evaluasi Model



Gambar 9. Grafik Perbandingan LSTM, SARIMAX dan Aktual

Berdasarkan gambar 9 prediksi curah hujan terlihat jelas, model LSTM secara konsisten memberikan hasil prediksi yang lebih mendekati data aktual dibandingkan model SARIMAX. Dengan hasil evaluasi MAE, RMSE dan R^2 ditunjukkan pada tabel 1

TABEL 1. NILAI EVALUSI MODEL

Evaluasi Model	SARIMAX	LSTM
RMSE	31.5	19.6
MAE	24.5	15.0
R ²	0,52	0,81

RMSE mengukur rata-rata besarnya kesalahan prediksi, dengan nilai RMSE SARIMAX adalah 31.5 sedangkan LSTM adalah 19.6. Nilai RMSE model LSTM memiliki nilai yang lebih rendah dari SARIMAX secara umum model yang dihasilkan LSTM lebih baik dari SARIMAX.

MAE mengukur nilai rata-rata besarnya kesalahan tetapi tanpa mengkuadratkan, MAE lebih mudah diinterpretasikan karena berada dalam skala yang sama dengan data aslinya. Nilai MAE SARIMAX adalah 24.5 dan nilai MAE LSTM adalah 15.0. Di lihat dari nilai MAE model LSTM lebih unggul. Artinya rata-rata selisih absolut antara nilai prediksi dan nilai aktual pada LSTM lebih kecil.

R² mengukur seberapa besar prosentase variasi dari data aktual yang dapat dijelaskan oleh model, nilai R² antara 0 hingga 1. Semakin mendekati angka 1 model semakin baik. Nilai evaluasi R² model SARIMAX adalah 0.52 sedangkan untuk LSTM nilainya 0.81. Model LSTM dapat menjelaskan sekitar 81% variasi data, hal ini menunjukkan tingkat kecocokan yang sangat baik. Sedangkan model SARIMAX mampu menjelaskan sekitar 52 % variasi pada data yang tergolong performa sedang.

IV. KESIMPULAN

Model LSTM menunjukkan kinerja yang lebih unggul dibandingkan dengan model SARIMAX dalam memprediksi curah hujan dasarian. Keunggulan LSTM terbukti secara kuantitatif dengan nilai RMSE yang lebih rendah 19.6 dibandingkan SARIMAX sebesar 31.5 serta MAE yang lebih kecil yaitu 15.0 dibandingkan SARIMAX 24.5. Nilai RMSE dan MAE yang lebih rendah menunjukkan tingkat kesalahan prediksi model LSTM lebih kecil. LSTM mampu menjelaskan 81% variabilitas data curah hujan aktual di Kota Semarang, yang jauh lebih baik dibandingkan SARIMAX yang hanya 52%. Hal ini menunjukkan LSTM jauh lebih baik dalam menangkap pola musiman dan trend yang kompleks serta lebih dinamis dan responsive terhadap lonjakan curah hujan ekstrem. Hasil penelitian ini dapat menjadi pertimbangan bagi instansi terkait dalam penanggulangan bencana hidrometeorologi (banjir, kekeringan, longsor dan lain-lain). Dukungan untuk sektor pertanian dalam merancang jadwal tanam sesuai dengan kebutuhan air bagi tanaman. Eksplorasi dan optimalisasi model LSTM lebih lanjut mengingat bahwa model LSTM terbukti lebih unggul dalam menjelaskan 81 % variabilitas data curah hujan dan memiliki tingkat kesalahan prediksi yang lebih kecil (RMSE 19,6, MAE 15,0) dibandingkan SARIMAX. Penelitian selanjutnya dapat berfokus pada optimalisasi arsitektur LSTM, meliputi penyesuaian parameter seperti jumlah *epoch*, *batch size*. Dan dapat ditambah variabel lain misalnya data angin, penyinaran matahari untuk meningkatkan akurasi prediksi.

REFERENSI

- [1] D. S. Adetama, A. Fauzi, B. Juanda, and D. B. Hakim, "Evaluasi Pembangunan Berkelanjutan dengan Rendah Karbon pada Sektor Pertanian Padi," *Tataloka*, vol. 25, no. 1, pp. 50–69, 2023, doi: 10.14710/tataloka.25.1.50-69.
- [2] A. Permahadi and E. Widowati, "Mitigasi bencana banjir di kota semarang," *Higeia J. Public Heal. Res. Dev.*, vol. 6, no. 2, pp. 225–235, 2022, [Online]. Available: <http://journal.unnes.ac.id/sju/index.php/higeia>
- [3] H. A. Ramadhan and I. V. Paputungan, "Prediksi Curah Hujan Menggunakan Metode Arima,"

- Edusaintek J. Pendidikan, Sains dan Teknol.*, vol. 12, no. 1, pp. 314–328, 2025.
- [4] H. R. AA Maulana, “IMPLEMENTASI METODE SARIMAX UNTUK PREDIKSI CURAH HUJAN JANGKA PENDEK DI PAGERAGEUNG, TASIKMALAYA,” *J. Sumber Daya Air*, vol. 20, no. 1, pp. 39–50, 2024, doi: 10.32679/jsda.v20i1.874.
 - [5] A. A. Akbar, Y. Darmawan, A. Wibowo, and H. K. Rahmat, “Accuracy Assessment of Monthly Rainfall Predictions using Seasonal ARIMA and Long Short-Term Memory (LSTM),” *JCSE J. Comput. Sci. an Eng.*, vol. 5, no. 2, pp. 99–114, 2024, [Online]. Available: <http://icsejournal.com/index.php/>
 - [6] E. Wati, Y. Arman, and R. Adriat, “KETAPANG BAGIAN SELATAN MENGGUNAKAN METODE AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA),” vol. 10, no. 3, pp. 323–331, 2022.
 - [7] N. M. M. Candra Devi, I. P. A. Bayupati, and N. K. A. Wirdiani, “Prediksi Curah Hujan Dasarian dengan Metode Vanilla RNN dan LSTM untuk Menentukan Awal Musim Hujan dan Kemarau,” *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 405, 2022, doi: 10.26418/jp.v8i3.56606.
 - [8] M. Z. Arkananta, Y. Darmawan, M. E. Trianasari, W. Widodo, and A. Widodo, “Perbandingan Model Prediksi Curah Hujan Kota Semarang Metode SARIMA Perbandingan Model Prediksi Curah Hujan Kota Semarang Metode SARIMA dan Holt-Winter,” no. July, 2025, doi: 10.29313/jrs.v5i1.6149.
 - [9] A. Wijayanto, A. Sugiharto, and R. Santoso, “Identifikasi Dini Curah Hujan Berpotensi Banjir Menggunakan Algoritma Long Short-Term Memory (Lstm) Dan Isolation Forest,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 3, pp. 637–646, 2024, doi: 10.25126/jtiik.938718.
 - [10] P. F. Nurhayati and S. Subekti, “Kebijakan Pemerintah dalam Mengatasi Banjir di Kota Semarang,” vol. 3, no. 1, pp. 85–92, 2022.
 - [11] C. P. Chai, “The Importance of Data Cleaning: Three Visualization Examples,” *Chance*, vol. 33, no. 1, pp. 4–9, 2020, doi: 10.1080/09332480.2020.1726112.
 - [12] B. K. Iwana and S. Uchida, *An empirical survey of data augmentation for time series classification with neural networks*, vol. 16, no. 7 July. 2021. doi: 10.1371/journal.pone.0254841.
 - [13] M. R. A. Yudianto, T. Agustin, R. M. James, F. I. Rahma, A. Rahim, and E. Utami, “Rainfall Forecasting to Recommend Crops Varieties Using Moving Average and Naive Bayes Methods,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 13, no. 3, pp. 23–33, 2021, doi: 10.5815/IJMECS.2021.03.03.
 - [14] AritaWitanti, “Smoothing Data Fluktuatif Dengan Exponential Smoothing Studi Kasus Data Curah Hujan,” *J. InFact*, vol. 1, no. 2, pp. 1–7, 2016.
 - [15] S. Maesaroh, K. Muludi, and J. Triloka, “Estimasi Jarak Pandang Meteorologi di Bandar Udara Menggunakan Metode Back Propagation dan CNN,” vol. 7, no. 1, pp. 1–9, 2025, doi: 10.47065/bits.v7i1.7138.
 - [16] H. Li, S. Li, and H. Ghorbani, “Data-driven novel deep learning applications for the prediction of rainfall using meteorological data,” *Front. Environ. Sci.*, vol. 12, no. August, pp. 1–15, 2024, doi: 10.3389/fenvs.2024.1445967.
 - [17] C. Chen *et al.*, “Forecast of rainfall distribution based on fixed sliding window long short-term memory,” *Eng. Appl. Comput. Fluid Mech.*, vol. 16, no. 1, pp. 248–261, 2022, doi: 10.1080/19942060.2021.2009374.
 - [18] P. Chuasuk, T. Bhatrasataponkul, and A. Akkarapongtrakul, “Comparative analysis and enhancing rainfall prediction models for monthly rainfall prediction in the Eastern Thailand,” *MethodsX*, vol. 14, no. December 2024, 2025, doi: 10.1016/j.mex.2024.103094.
 - [19] M. Z. Bin Hilmi, T. Anwar, and D. R. B. A. Rambli, “Long Short-Term Memory With Gated Recurrent Unit Based on Hyperparameter Settings and Hybridization for Reference Evapotranspiration Rate Prediction,” *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 22, pp. 6702–6714, 2022.
 - [20] M. Chhetri, S. Kumar, P. P. Roy, and B. G. Kim, “Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan,” *Remote Sens.*, vol. 12, no. 19, pp. 1–13, 2020, doi: 10.3390/rs12193174.
 - [21] X. Wu *et al.*, “The development of a hybrid wavelet-arima-lstm model for precipitation amounts and drought analysis,” *Atmosphere (Basel)*, vol. 12, no. 1, pp. 1–17, 2021, doi: 10.3390/ATMOS12010074.