172 ANALISIS EFEKTIVITAS TEKNIK IMPUTASI PADA LSTM UNTUK MENINGKATKAN KUALITAS DATA PADA PERAMALAN CURAH HUJAN

By Ariyanto Adi Nugroho

WORD COUNT 3691 TIME SUBMITTED 15-MAY-2024 12:24PM

PAPER ID 108975866

Template ini berlaku untuk terbitan Volume 6 Nomor 1 April 2023

ANALISIS EFEKTIVITAS TEKNIK IMPUTASI PADA LSTM UNTUK MENINGKATKAN KUALITAS DATA PADA PERAMALAN CURAH HUJAN

Ariyanto Adi Nugroho

[12,] Program Studi Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Nusa Mandiri, Jakarta

Nusa Mandiri Tower Jl. Jatiwaringin Raya No. 2, Jakarta Timur 13620
[1] 14210241@nusamandiri.ac.id, [2] muhammad.uhs@nusamandiri.ac.id

Abstract

Climate monitoring data obtained from meteorological stations can have missing values due to various reasons. Data incompleteness occurs because of transmission failure, non-responsive sensors, equipment repairs, and other issues. The problem triggers inconsistent data and noise in climate measurement data. A better solution to handle missing values in weather data proposed on this work. Data imputation method address these issues before further analysis is conducted. This research proposes the application of imputation techniques during the data preparation phase. The research findings indicate that the best imputation method is KNN combined with Bidirectional LSTM. The evaluation metric results are Mean Absolute Error (MAE) 3,3599, Mean Square Error (MSE) 78,4336, Root Mean Squared Error (RMSE) 8,8562 and R-Square 0,5365.

Keywords : LSTM, LOCF, KNN, NOCB, MICE

Abstrak

Data pemantauan iklim yang didapatkan dari stasiun meteorologi dapat memiliki *missing value* karena berbagai hal. Ketidaklengkapan data dapat terjadi karena transmisi gagal, sensor tidak merespons, perbaikan perangkat, dan lain-lain. Masalah yang didapati umumnya adalah data tidak konsisten dan adanya *noise* pengukuran data iklim. Diperlukan solusi penanganan *missing values* pada data cuaca agar dapat diatasi sebelum dilakukan analisis lebih lanjut. Penelitian ini mengusulkan penerapan *data imputation* pada fase *data preparation* menyesuaikan karakteristik data. Metode *forecasting* yang diterapkan adalah LSTM dan Bidirectional LSTM yang merupakan turunan dari RNN. Metode ini menghasilkan model dari data *time series* yang lebih baik dibanding RNN. Hasil penelitian menyimpulkan metode imputasi yang memiliki performa terbaik adalah KNN dipadukan dengan metode Bidirectional LSTM. Nilai evaluation metric yang diperoleh adalah Mean Absolute Error (MAE) sebesar 3,3599, Mean Square Error (MSE) sebesar 78,4336, Root Mean Squared Error (RMSE) sebesar 8,8562 dan R-Squared sebesar 0,5365.

Kata kunci : LSTM, LOCF, KNN, NOCB, MICE

1. PENDAHULUAN

Data berperan penting dalam menentukan keberhasilan hasil penelitian. Isu utama yang sering muncul dalam kualitas data adalah keberadaan nilai yang hilang atau '*missing values*'. Masalah ini timbul ketika ada bagian dari data yang tidak tersedia atau hilang. Fenomena nilai hilang sering terjadi dalam kumpulan data, yang disebabkan oleh berbagai faktor seperti kerusakan perangkat, kesalahan perhitungan, kegagalan pencatatan data, serta masalah teknis lainnya [1]. Nilai yang hilang ini seringkali menjadi penghalang. Data yang hilang biasanya penting. Hilangnya nilai dapat mengakibatkan proses analisis menjadi tidak akurat, tidak efisien dan menurunkan akurasi [2].

Mekanisme *missing values* atau nilai yang hilang menurut Donald B. Rubin dikelompokkan menjadi tiga [3]. Mekanisme pertama adalah *Missing at Completely Random* (MCAR). MCAR terjadi ketika *missing value* tidak memiliki relasi atau dependensi pada data yang diobservasi, tidak diobservasi maupun missing data itu sendiri [3]. Mekanisme kedua adalah *Missing at Random* (MAR). MAR terjadi ketika *missing value* memiliki relasi dengan nilai yang diobservasi [3]. MCAR mengacu pada situasi di mana kehilangan data terjadi secara sepenuhnya acak, tanpa ada keterkaitan dengan data atau atribut lain. Pada MAR, kehilangan data berkaitan dengan atribut lain atau dengan data yang teramati. Contoh MAR adalah responden perempuan cenderung tidak berkenan menyebutkan umur maupun berat badan. Kehilangan data pada umur dan berat badan berhubungan dengan jenis kelamin. Mekanisme ketiga adalah *Missing Not At Random* (MNAR). MNAR terjadi ketika *missing value* memiliki relasi dengan nilai yang diobservasi. MNAR terjadi ketika kehilangan data berkaitan langsung dengan nilai yang hilang itu sendiri, dimana nilai yang tidak tercatat berhubungan dengan suatu kejadian yang tidak teramati.[1] [4].

Missing value juga sering ditemukan pada data berjenis *time series*. *Time series* merupakan rentetan data yang dikumpulkan secara berkala dan disusun berdasarkan urutan waktu [5]. Metode statistik awalnya digunakan untuk melakukan prediksi *time series*. Pada perkembangannya *deep learning* dengan metode *regression* pada akhirnya mengungguli metode statistik yang sudah diterapkan dari tahun 1970 [6]. Model *deep learning* didapati dapat menganalisis data *time series* secara terukur dan akurat [7]. *Deep learning model* dikenal lazim diterapkan pada data sekuensial atau *time series*. Metode-metode yang termasuk adalah *Convolutional Neural Network* (CNN), *Gated Recurrent Units (GRU), Recurrent Neural Networks* (RNN), *Long Short-Term Memory* (LSTM) [8].

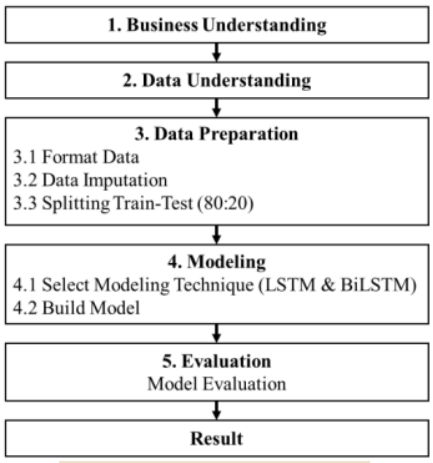
LSTM didesain untuk memecahkan isu *vanishing gradient* pada RNN [9]. Isu ini terjadi ketika nilai *gradient* terlalu kecil sehingga model berhenti melakukan *training* [10]. Model LSTM memiliki kemampuan yang baik untuk mengingat dependensi jangka panjang karena memiliki *constructed cell* untuk menyimpan informasi. Prinsip dari *cell* ini untuk mengelola pembaruan memori jangka sehingga informasi dan *gradient* dapat mengalir tanpa ada perubahan selama iterasi berlangsung [8].

Data pemantauan iklim yang didapatkan dari stasiun meteorologi dapat memiliki *missing value* karena berbagai hal. Ketidaklengkapan data dapat terjadi karena transmisi gagal, sensor tidak merespons, perbaikan perangkat, dan lain-lain. Masalah yang didapati umumnya adalah data tidak konsisten dan adanya *noise* pengukuran data iklim. Diperlukan solusi penanganan *missing values* pada data cuaca agar dapat diatasi sebelum dilakukan analisis lebih lanjut [11]. Hal ini penting karena *dataset* yang lengkap dapat mempengaruhi pengambilan keputusan yang memanfaatkan data tersebut. Data dengan kualitas yang rendah dapat menyebabkan analisis yang tidak akurat yang berujung kesalahan pengambilan kebijakan [3].

Melihat dari pemaparan diatas, penelitian dilakukan untuk menerapkan variasi metode *data imputation* dan algoritma LSTM yang terbaik. Penelitian dilakukan dalam rangka meningkatkan kualitas data cuaca. Data yang telah melalui *data imputation* kemudian dijadikan *training data*. Model kemudian dibangun dari *training data* yang telah siap. Model kemudian diuji menggunakan *testing data.* Selanjutnya dilakukan perbandingan antara kombinasi metode *data imputation* dan algoritma LSTM dan Bidirectional LSTM (Bi LSTM). Penelitian kemudian dilanjutkan dengan menghitung nilai *evaluation metric* untuk mengukur performa model yang dihasilkan. Data asli dan prediksi kemudian dibandingkan dan divisualisasikan.

2. METODOLOGI PENELITIAN

Penelitian mengadopsi metode Cross Industry Standard Process for Data Mining (CRISP-DM) dari tahapan Business Understanding hingga Evaluation. Metodologi penelitian yang digunakan dan diterapkan diilustrasikan pada Gambar 1 dibawah.



Gambar 1 Metodologi Penelitian

1. *Business Understanding*

Fase *business understanding* digunakan tujuan dan sasaran dari penelitian dipahami dan didefinisikan. Penelitian dilakukan untuk

menentukan metode data imputation terbaik untuk meningkatkan *forecast* data iklim dari Stasiun Meteorologi Kelas III Kemayoran Jakarta Pusat selama 3.530 hari dari Januari 2011 hingga Agustus 2020. Penelitian dilakukan *multivariate* pada parameter curah hujan dengan nama *field* Tavg, RH_avg dan RR.

2. Data Understanding

Dataset yang diunduh berasal dari Stasiun Meteorologi Kemayoran adalah Unit Pelaksana Teknis (UPT) yang berkedudukan pada kantor pusat BMKG.

Data stasiun kemayoran terdiri atas 11 *field* yang terdiri atas Tanggal, Tn (suhu minimum dalam derajat Celcius), Tx (suhu maksimum dalam derajat Celcius), Tavg (suhu rata-rata dalam derajat Celcius), RH_avg (pengukuran kelembaban rata-rata dengan satuan %), RR (curah hujan dengan satuan mm), ss (lamanya durasi penyinaran matahari dengan satuan jam), ff_x (pengukuran kecepatan angin maksimum dengan satuan m/s), ddd_x (arah angin terdeteksi saat kecepatan maksimum dengan satuan °) , ff_avg (pengukuran kecepatan angin rata-rata dengan kecepatan m/s), dan ddd_car (arah angin terbanyak yang tercatat dengan satuan °).

Eksperimen dilakukan secara multivariate dengan target field Tavg, RH_avg dan RR. Field dipilih karena Tavg menunjukkan variabilitas yang sesuai. Kelembaban relatif (RH_avg) memiliki rata-rata 75,73% dengan standar deviasi 6,27% yang menunjukkan variabilitas tetapi masih dalam kisaran yang umum untuk iklim Indonesia yang cenderung lembab. Curah hujan (RR) memiliki rentang yang luas dari 0 hingga 277,5 mm dengan rata-rata 6,29 mm.

3. Data Preparation

Dataset diunduh pada rentang Januari 2011 sampai Agustus 2020 dari aplikasi pada URL https://dataonline.bmkg.go.id/home. Aplikasi ini disediakan oleh BMKG untuk menyediakan layanan data bagi kalangan internal maupun eksternal yang terdiri dari perguruan tinggi, institusi kementerian, lembaga, swasta dan masyarakat. Penulis tidak melakukan pengukuran dengan sensor mandiri seperti pada tulisan [12] dan Handayani [13].

Dataset diunduh pada rentang Januari 2011 sampai Agustus 2020 dari aplikasi pada URL https://dataonline.bmkg.go.id/home. Aplikasi ini disediakan oleh BMKG untuk menyediakan layanan data bagi kalangan internal maupun eksternal yang terdiri dari perguruan tinggi, institusi kementerian, lembaga, swasta dan masyarakat. Langkah untuk akuisisi data adalah sebagai berikut:

1. Cek ketersediaan data iklim pada menu "Ketersediaan Data"
2. Bagi pengguna baru, klik menu Registrasi dan isi formulir registrasi dengan benar. Aktivasi akun dengan mengakses link verifikasi pada alamat e-mail yang didaftarkan
3. Bagi pengguna dengan akun aktif, dapat login dengan email dan password yang teraktivasi. Isi kode Captcha untuk menyelesaikan proses login
4. Pilih menu Data Iklim, lalu pilih menu Data Harian
5. Pilih Jenis Stasiun, Parameter, Provinsi, Kabupaten, No/Nama Stasiun dan Rentang Waktu. Isian yang digunakan adalah sebagai berikut:
   - Jenis Stasiun: UPT
   - Parameter: arah angin saat kecepatan maksimum (ddd_x), arah angin terbanyak (ddd_car), curah hujan (RR), kecepatan angin maksimum (ff_x), kecepatan angin rata-rata (ff_avg), kelembaban rata-rata (RH_avg), lamanya penyinaran matahari (ss), tempratur maksimum (Tx), tempratur minimum (Tn), tempratur rata-rata (Tavg)
   - Provinsi: DKI Jakarta
   - Kabupaten: Kota Adm. Jakarta Pusat
   - No/Nama Stasiun: 96745 Stasiun Meteorologi Kemayoran
   - Rentang Waktu: hari pertama bulan dan hari terakhir bulan yang akan diunduh
6. Klik tombol Proses. Aplikasi kemudian menampilkan kolom penilaian pelayanan. Isi penilaian pelayanan dan data akan terunduh.
7. Klik nama Profil (misalnya Ariyanto Adi Nugroho) dan klik Logout untuk keluar dari aplikasi

Data diunduh dalam format Microsoft Excel dengan ekstensi .xlsx. Satu file Microsoft Excel berisi data pengukuran selama 1 bulan. Terdapat 116 file yang diunduh untuk mendapatkan data pengukuran dari Januari 2011 hingga Agustus 2020.

Data diformat untuk mempersiapkannya dengan beberapa tahapan, antara lain skiprows, drop, to_datetime, replace, dan append. Seluruh

parameter menggunakan library Pandas. Pandas adalah library bahasa pemrograman Python untuk melakukan manipulasi dan analisis terhadap string, numeric, datetime dan time series data [29]. Penjelasannya adalah sebagai berikut:

a. skiprows

Parameter pada function pandas.read_excel(). Parameter ini dapat digunakan untuk melewati sejumlah baris pada saat awal pembacaan file untuk dapat diubah menjadi DataFrame. Tipe isian data yang diterima dapat berupa integer. Penggunaan parameter ini dalam rangka mengabaikan header dari pada file dataset.

b. drop

Function pada DataFrame pandas. Function digunakan untuk menghilangkan sejumlah baris atau kolom pada DataFrame. Tipe isian data yang diterima dapat berupa string atau daftar beberapa string. Penggunaan parameter ini dalam rangka menghilangkan data yang tidak relevan agar peneliti dapat fokus pada data yang relevan.

c. to_datetime

Function pada pandas untuk melakukan konversi format data menjadi datetime. Tipe isian data yang diterima dapat berupa string, angka maupun list berisi tanggal dan waktu pada parameter 'arg', format tanggal waktu pada 'format', dan konstanta ignore, raise, dan coerce pada parameter 'errors'. Penggunaan parameter ini dalam rangka standarisasi format tanggal dan waktu.

d. replace

Function pada pandas untuk melakukan penggantian nilai pada DataFrame. Penggunaan parameter ini dalam rangka membersihkan, mempersiapkan data, dan mengganti nilai yang tidak valid seperti 8888 dan 9999 yang ditemui pada dataset.

e. append

Function pada pandas untuk menambahkan baris pada DataFrame. Penggunaan parameter ini dalam rangka menggabungkan dataset pengukuran data iklim selama 116 bulan dari Januari 2011 sampai Agustus 2020.

Langkah terakhir dilakukan data imputation dengan metode yang sesuai. Data kemudian dipisah menjadi 80% testing data dan 20% training data.

4. Modeling

Fase modeling dilakukan dengan melakukan penerapan LSTM dan Bi LSTM untuk melakukan training menggunakan TensorFlow dan Keras. Percobaan pertama diterapkan menggunakan metode LSTM. Percobaan kedua diterapkan dengan Bidirectional LSTM.

5. Evaluation

Pada fase evaluation performa model diukur. Pengukuran performa menggunakan Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) dan R-Square.

a. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) merupakan nilai pengukuran error antara observasi yang disandingkan yang memiliki fenomena yang sama. MAE mengukur kesalahan rata-rata absolut dalam prediksi statistik at[3] machine learning. Metrik pengukuran MAE menghitung rata-rata dari selisih absolut antara nilai yang diprediksi dan nilai sebenarnya [30]. Makin kecil nilai MAE menunjukkan perbedaan antara nilai prediksi dan nilai aktual minimal. Dapat disimpulkan model lebih akurat dalam melakukan prediksi [31]. Rumus perhitungan MAE adalah sebagai berikut:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Nilai n adalah jumlah sampel, yi adalah nilai sebenarnya dan $\hat{y}_i$ adalah nilai prediksi [32].

b. Mean Squared Error (MSE)

Mean Squared Error (MSE) adalah[3] metrik pengukuran kualitas estimator, yakni rata-rata dari kuadrat kesalahan atau perbedaan antara nilai yang diestimasi dan nilai sebenar[5]a [32]. Secara matematis, MSE dihitung sebagai rata-rata dari kuadrat perbedaan antara nilai prediksi yang dihasilkan oleh model dan nilai aktual. Nilai MSE yang rendah menunjukkan bahwa model memiliki kesalahan prediksi yang kecil. Kesalahan prediksi yang kecil maknanya kinerjanya baik dalam memprediksi hasil. Rumus perhitungan MSE adalah sebagai berikut:

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

Nilai n adalah jumlah s[10]pel, *yi* adalah nilai sebenarnya untuk sampel ke-i dan $\hat{y}_i$ adalah nilai prediksi untuk sampel ke-i [32].

c. Roo[3]Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) merupakan metrik evaluasi yang digunakan untuk mengukur rata-rata kesalahan kuadrat antara nilai-nilai yang diobservasi (atau nilai sebenar[8]a) dan nilai-nilai yang diprediksi oleh model. RMSE merupakan akar kuadrat dari Mean Squared Error (MSE). RMSE memberikan estimasi ukuran kesalahan model dalam unit yang sama dengan variabel yang diprediksi. Rumus perhitungan RMSE adalah sebagai berikut:

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

Nilai n adalah jumlah sampel, *yi* adalah nilai sebenarnya untuk sampel ke-i dan $\hat{y}_i$ adalah nilai prediksi untuk sampel ke-i [33].

d. R-Squared

R-squared atau koefisien determinasi, adalah metrik statistik yang digunakan untuk mengukur proporsi variansi dalam variabel dependen yang dapat dijelaskan oleh variabel independen dalam model regresi. R-squared adalah ukuran seberapa baik prediksi model sesuai dengan data aktual. Nilai R-Squared dinyatakan pada rentang 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa model dapat menjelaskan proporsi variansi yang lebih besar dalam variabel dependen.

Nilai R-Square dihitung dengan rumus nilai 1 dikurangi hasil bagi dari SSres dan SStot. SSres adalah jumlah kuadrat residu (sum of squared residuals) yang mengukur variasi antara nilai yang diamati dan nilai yang diprediksi oleh model. SStot adalah jumlah kuadrat total (total sum of squares) yang mengukur variasi total dari nilai yang diamati [34]. Perhitungan R-Square dapat dituliskan sebagai persamaan sebagai berikut:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

6. Deployment

Penelitian ini tidak menerapkan fase Deployment dari CRISP-DM.

3. HASIL DAN PEMBAHASAN

Eksperimen dilakukan secara *multivariate* dengan *target field* Tavg, RH_avg dan RR. *Field* dipilih karena Tavg menunjukkan variabilitas yang sesuai. Kelembaban relatif (RH_avg) memiliki rata-rata 75,73% dengan standar deviasi 6,27% yang menunjukkan variabilitas tetapi masih dalam kisaran yang diharapkan untuk iklim Indonesia yang cenderung lembab. Curah hujan (RR) memiliki rentang yang luas dari 0 hingga 277,5 mm dengan rata-rata 6,29 mm.

Dataset yang dipublikasikan berformat Microsoft Excel. Satu file Microsoft Excel berisi data pengukuran selama 1 bulan. Terdapat 116 file yang diunduh untuk mendapatkan data pengukuran dari Januari 2011 hingga Agustus 2020.

Dataset yang telah diperoleh kemudian melalui tahap *data preparation* (penghapusan teks yang tidak diperlukan, penghapusan *noise*, dan penerapan *data imputation*).

Eksperimen dilakukan sebanyak 16 kali yang dibagi menjadi 2 kelompok. Kelompok pertama menerapkan LSTM dan kelompok kedua menerapkan Bidirectional LSTM. Skema pembagiannya dijelaskan pada Tabel 1.

Tabel 1 Metode *Data Imputation* dan Metode *Forecasting*

Eksperimen	Metode *Data Imputation*	Metode *Forecasting*
LSTM.1	-	LSTM
LSTM.2	Mean	
LSTM.3	Median	
LSTM.4	Mode	
LSTM.5	LOCF	
LSTM.6	NOCB	
LSTM.7	kNN	
LSTM.8	MICE	
BiLSTM.9	-	Bi LSTM
BiLSTM.10	Mean	
BiLSTM.11	Median	
BiLSTM.12	Mode	
BiLSTM.13	LOCF	
BiLSTM.14	NOCB	
BiLSTM.15	kNN	
BiLSTM.16	MICE	

Tabel 2 Parameter dan Nilai Standar

Nama Parameter	Nilai Standar pada Eksperimen
Window Size	240
Dense Units	25 neurons
Dropout Rate	0,5 (50%)
Optimizer	Adam optimizer
Number of Neurons in Dense Layer	25
Epochs	50 epoch
Batch Size	500
Loss Function	Mean Squared Error

Hyperparameter pada 16 eksperimen pada Tabel 1 didaftar pada Tabel 2. Penjelasan masing-masing *hyperparameter* adalah sebagai berikut:

a. *Window Size,* ukuran yang menentukan jumlah *time steps* sebelumnya yang akan digunakan untuk memprediksi nilai saat ini. Setiap sample input akan terdiri dari 240 *time step* sebelumnya. Ukuran *Window Size* penting karena menentukan seberapa jauh mundur kebelakang model harus melihat untuk membuat prediksi.
b. *Dense Units*, jumlah unit (neuron) dalam lapisan LSTM. Jumlah neuron di setiap lapisan LSTM maupun dalam lapisan Bi-LSTM ditentukan oleh nilai ini. Jumlah neuron menentukan kapasitas model untuk

menangkap informasi dan kompleksitas dalam data.

c. *Dropout Rate*, adalah proporsi neuron yang dinonaktifkan secara acak selama proses pelatihan untuk mencegah terjadinya *overfitting*. Nilai dropout rate ada pada rentang 0 dan 1.
d. *Optimizer* adalah algoritma optimasi yang digunakan untuk meminimalkan *loss fuction.*
e. *Number of Neurons in Dense Layer*, merupakan jumlah neuron pada lapisan dense. Digunakan untuk pemrosesan lebih lanjut setelah ekstraksi fitur oleh lapisan LSTM dan Bi-LSTM sebelum output akhir diproduksi
f. *Epochs* adalah jumlah lengkap iterasi pelatihan model atas keseluruhan dataset
g. Batch Size adalah jumlah sampel data yang diproses sebelum model melakukan pembaruan bobot. Jika nilai *batch size* adalah 32 maka 32 sampel akan diproses sebelum melakukan satu kali pembaruan bobot. Angka *batch size* harus mampu menyeimbangkan antara efisiensi komputasi dan keakuratan estimasi gradien.

1. *Loss Function* atau fungsi kerugian yang menghitung rata-rata kuadrat perbedaan antara nilai yang diprediksi oleh model dan nilai sebenarnya. *Loss function* akan memberikan gambaran seberapa baik model melakukan prediksi.

1. Eksperimen Menggunakan Metode LSTM

Kelompok eksperimen dilakukan pertama kali terhadap dataset menggunakan metode LSTM dengan hasil pada Tabel 3

Nilai rujukan yang digunakan adalah MAE 11,3 yang dihasilkan oleh penelitian "Prediksi Curah Hujan Harian di Stasiun Meteorologi Kemayoran menggunakan Artificial Neural Network (ANN)" yang ditulis oleh Richard Mahendra Putra dan Nurhastuti Anjar Rani. Paper tersebut menggunakan rentang dataset yang sama yaitu Januari 2011 hingga Agustus 2020 [14]. Tabel 3 memaparkan 8 eksperimen yang menggunakan data imputation dan menerapkan LSTM. Hasil yang diperoleh secara umum memiliki performa MAE dibawah nilai 11,3.

Tabel 3 Hasil Eksperimen Metode LSTM

Eksperimen	Metode Forecasting	Metode Data Imputation	Evaluation Metrics			
			MAE	MSE	RMSE	R-Square
Rujukan	ANN	-	11,3	-	-	-
LSTM.1	LSTM	-	3,9079	131,0827	11,4491	0,3272
LSTM.2	LSTM	Mean	3,6772	115,2757	10,7366	0,3422
LSTM.3	LSTM	Median	3,6715	119,4415	10,9289	0,3306
LSTM.4	LSTM	Mode	3,5651	116,8508	10,8097	0,3503
LSTM.5	LSTM	LOCF	3,7328	124,4447	11,1554	0,3311
LSTM.6	LSTM	NOCB	3,8043	125,7679	11,2146	0,3441
LSTM.7	LSTM	KNN	3,7243	121,9033	11,0409	0,3262
LSTM.8	**LSTM**	**MICE**	**3,6319**	**116,3358**	**10,7859**	**0,3582**

Eksperimen menghasilkan nilai MAE dibawah 11,3. Nilai MAE terendah didapat pada Eksperimen LSTM.4 dengan nilai 3,5651. Nilai MSE dan RMSE terbaik didapat pada Eksperimen 2 yang mampu menghasilkan prediksi yang mendekati actual value dan memiliki tingkat error yang rendah. Nilai R-Squared terbaik diperoleh pada Eksperimen LSTM.8 dengan nilai 0,358 yang berarti memiliki proporsi varians tertinggi pada variabel dependen yang dapat diprediksi dari variabel independen.

Eksperimen LSTM.8 dapat disimpulkan sebagai eksperimen terbaik pada kelompok ini karena nilai yang didapat merepresentasikan keseimbangan akurasi dan kemampuan model menjelaskan varian (*explain the variance*).

2. Eksperimen Menggunakan Metode Bidirectional LSTM (Bi LSTM)

Kelompok eksperimen kedua dilakukan terhadap dataset dengan menggunakan Bi LSTM dengan hasil pada Tabel 4 menjabarkan 8 eksperimen yang menerapkan data imputation dan menggunakan metode Bi LSTM.

Tabel 4 Hasil Eksperimen Metode Bidirectional LSTM (Bi LSTM)

Eksperimen	Metode *Forecasting*	Metode *Data Imputation*	Evaluation Metrics			
			MAE	MSE	RMSE	R-Square
Rujukan	ANN	-	11,3	-	-	-
BiLSTM.1	LSTM	-	4,8895	179,8081	13,4092	0,1284
BiLSTM.2	LSTM	Mean	4,0619	135,2290	11,6288	0,2570
BiLSTM.3	LSTM	Median	4,1676	135,5840	11,6440	0,2660
BiLSTM.4	LSTM	Mode	4,0849	132,0755	11,4924	0,2481
BiLSTM.5	LSTM	LOCF	3,4059	85,2731	9,2343	0,5282
BiLSTM.6	LSTM	NOCB	3,3998	81,6515	9,0361	0,5430
BiLSTM.7	**LSTM**	**KNN**	**3,3599**	**78,4336**	**8,8562**	**0,5365**
BiLSTM.8	LSTM	MICE	3,4103	80,0259	8,9457	0,5271

Penerapan metode *data imputation* mean, median, dan mode menampilkan hasil yang tidak jauh berbeda dengan eksperimen menggunakan metode LSTM. Hasil eksperimen menampilkan hasil yang bervariasi. Nilai MAE terendah didapat pada Eksperimen BiLSTM.7 dengan nilai 3,3599. Nilai MSE dan RMSE terbaik didapat pada Eksperimen BiLSTM.7 dengan MSE 78.4336 dan RMSE 8.8562. Eksperimen BiLSTM.7 mampu menghasilkan prediksi yang mendekati actual value dan memiliki tingkat error yang rendah. Nilai R-Squared terbaik diperoleh pada Eksperimen BiLSTM.6 dengan nilai 0.543. Nilai R-Square ini menu berarti model memiliki kekuatan penjelasan terbaik, mampu menjelaskan proporsi yang lebih besar dari varians pada variabel dependen.

Pada eksperimen-eksperimen diatas, terlihat bahwa terdapat variasi yang signifikan dalam kinerja model antara delapan eksperimen berbeda. MSE, RMSE, dan MAE memberikan ukuran tentang kesalahan prediksi model. Ketiga metrik tersebut ketika menghasilkan nilai lebih rendah maka menunjukkan kinerja yang lebih baik.

Nilai R-Squared dibagi menjadi tiga klasifikasi: kuat, moderat, dan lemah, dengan nilai 0,75 dianggap kuat, 0,50 sebagai moderat, dan 0,25 sebagai lemah. R-Squared adalah evaluation metric yang melengkapi MSE untuk mengukur performa model yang berasal dari data timeseries [15]. R-Square yang diperoleh 0,5365 yang masuk pada klasifikasi moderat.

Dari hasil yang diperoleh, dapat disimpulkan Eksperimen BiLSTM.7 yang menggunakan metode *data imputation* KNN dan metode BiLSTM merupakan eksperimen dengan performa terbaik. *Evaluation metric* yang diperoleh adalah MAE 3,3599, MSE 78,4336, RMSE 8,8562, dan R-Square 0,5365. Hasil eksperimen BiLSTM.7 juga mengungguli Eksperimen LSTM.8 sebagai perbandingan.

Keunggulan hasil eksperimen BiLSTM.7 ditunjang penerapan *data imputation* KNN yang menerapkan *similarity-based imputation*. Metode KNN bekerja dengan menemukan tetangga terdekat dari titik data yang hilang dan memasukkan nilai berdasarkan kesamaan antar titik data. Mengingat sifat data cuaca yang kontinu dan dapat diprediksi, KNN dapat secara efektif memanfaatkan korelasi antara titik waktu terdekat atau kondisi serupa untuk memperhitungkan nilai yang hilang secara akurat.

Dataset cuaca memiliki *temporal correlation* atau korelasi temporal. Data cuaca pada dasarnya bersifat temporal dan kondisi dari satu titik waktu sering kali berkaitan erat dengan kondisi di titik waktu terdekat. KNN dapat memanfaatkan korelasi temporal ini, terutama jika kumpulan data disusun dalam *timeseries*, untuk menemukan pola imputasi yang serupa.

Pendekatan multivariat yang menargetkan field Tavg, RH_avg dan RR. Field ini saling terkait membuat KNN dapat menangani data multivariat dengan baik. KNN mempertimbangkan ruang multidimensi yang diciptakan oleh variabel-variabel ini, sehingga memungkinkan imputasi yang lebih akurat dengan mempertimbangkan hubungan antara berbagai variabel meteorologi.

4. Kesimpulan dan Saran

Eksperimen menggunakan dataset curah hujan BMKG yang telah melalui proses *data preprocessing* (skiprows, drop, to_datetime, replace, dan append) sebelum dilakukan *data imputation* dengan metode KNN. Eksperimen dengan kode BiLSTM.7 menghasilkan nilai MAE sebesar 3,3599, MSE sebesar 78,4336, RMSE sebesar 8,8562, dan R-Squared sebesar 0,5365.

Eksperimen BiLSTM.7 menunjukkan hasil yang paling baik dengan MAE, MSE, RMSE terendah, dan R-Square Score yang paling mendekati 1. Dengan demikian menandakan kesesuaian terbaik dengan data dan kesalahan prediksi paling kecil. Hasil-hasil tersebut menunjukkan Eksperimen BiLSTM.7 memiliki kinerja yang paling efektif dalam hal ketepatan prediksi dan kesesuaian model dengan dataset.

Metode data imputation terbukti meningkatkan hasil karena menerapkan *imputation* berbasis kesamaan. Dataset juga memiliki kecocokan dengan metode ini karena

bersifat temporal dan memiliki relasi antara *field* yang ditargetkan.

Sementara itu, eksperimen lain menunjukkan kinerja yang bervariasi dengan nilai R-Squared lebih rendah dan kesalahan prediksi yang lebih tinggi, yang mengindikasikan kesesuaian model yang kurang optimal dengan data.

Ada beberapa hal yang dapat dianalisis lebih lanjut untuk meningkatkan kegunaan dari hasil penelitian:

1. Penelitian selanjutnya dapat melakukan modifikasi *hyperparameter* agar menghasilkan model yang lebih sesuai
2. Pada penelitian selanjutnya dapat menerapkan *cross-validation* sehingga dapat memperoleh konfigurasi *hyperparameter* yang lebih optimal dibanding *default value* yang telah diterapkan.

172 ANALISIS EFEKTIVITAS TEKNIK IMPUTASI PADA LSTM UNTUK MENINGKATKAN KUALITAS DATA PADA PERAMALAN CURAH HUJAN

ORIGINALITY REPORT

12%
SIMILARITY INDEX

PRIMARY SOURCES

1	dataonline.bmkg.go.id Internet	142 words — 4%
2	Hendriyo Mokodompit, Nurnaningsih Nico Abdul, Elvie Fatmah Mokodongan. "PONDOK PESANTREN MODERN DARUL MADINAH WONOSARI KABUPATEN BOALEMO DENGAN PENDEKATAN ARSITEKTUR TROPIS", JAMBURA Journal of Architecture, 2024 Crossref	48 words — 1%
3	repository.unugiri.ac.id Internet	45 words — 1%
4	journal.stekom.ac.id Internet	21 words — 1%
5	repository.upi.edu Internet	21 words — 1%
6	wikistatistika.com Internet	20 words — < 1%
7	Richard Mahendra Putra, Nurhastuti Anjar Rani. "Prediksi Curah Hujan Harian di Stasiun	19 words — < 1%

Meteorologi Kemayoran Menggunakan Artificial Neural Network (ANN)", Buletin GAW Bariri, 2020
Crossref

8	www.scribd.com Internet	19 words — < 1%
9	dl.lib.uom.lk Internet	17 words — < 1%
10	repository.unja.ac.id Internet	16 words — < 1%
11	Widi Setiana, Della Andina, Nabhilah Deviani, Numan Musyaffa. "Implementasi Data Mining Untuk Analisa Data Penjualan Cat Menggunakan Algoritma Apriori dan Fp Growth (Studi Kasus PT.Sumbermas Unggul Nastari)", Jurnal Larik: Ladang Artikel Ilmu Komputer, 2021 Crossref	15 words — < 1%
12	teknosi.fti.unand.ac.id Internet	15 words — < 1%
13	www.researchsquare.com Internet	14 words — < 1%
14	dspace.library.uu.nl Internet	13 words — < 1%
15	core.ac.uk Internet	12 words — < 1%
16	repository.its.ac.id Internet	10 words — < 1%
17	ejurnal.itenas.ac.id Internet	9 words — < 1%

18	Wahyu Eko Cahyono, Dedy Kunhadi. "Strategi Pengembangan UKM Gethuk Pisang Guna Melestarikan Makanan Tradisional", Jurnal Media Teknik dan Sistem Industri, 2020 Crossref	8 words — < 1%
19	ejurnal.unikarta.ac.id Internet	8 words — < 1%
20	es.scribd.com Internet	8 words — < 1%
21	medicalclinic.id Internet	8 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF