JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989

DOI: 10.26418/justin.v12i2.76166
Vol. 12, No. 2, April 2024

Analisis Struktur Terbaik *Neural Network* dengan Algoritma *Backpropagation* dalam Memprediksi Indeks Kandungan Sulfida (SO_2) di Ibu Kota Jakarta

Dian Kurniasari[a,1], M Naufal Ammar Rafdiono[a,2], Warsono[a,3]

[a]*Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung, Indonesia*
Jl. Prof. Sumantri Brojonegoro No.1, Gedong Meneng, Bandar Lampung, 35145

[1]dian.kurniasari@fmipa.unila.ac.id
[2]naufalammar920@gmail.com
[3]warsono.1963@fmipa.unila.ac.id

Abstrak

Polusi udara merupakan masalah lingkungan yang umum terjadi di kota-kota besar di tanah air, tidak terkecuali Ibu Kota Jakarta. Tingginya jumlah penduduk di Jakarta menyebabkan konsentrasi polusi udara semakin tinggi karena peningkatan jumlah kendaraan. Kondisi ini diperburuk dengan banyaknya limbah yang dihasilkan oleh pabrik dari berbagai industri. Sulfida (SO_2) merupakan salah satu polutan dengan konsentrasi tertinggi di Jakarta dengan total beban emisi sebesar 19.7 kton. Oleh karena itu, prediksi indeks kandungan SO_2, merupakan isu penelitian yang penting karena zat SO_2 dapat berdampak terhadap berbagai faktor, seperti lingkungan, pertanian, dan kesehatan. Tujuan dari penelitian ini adalah menemukan model terbaik dalam melakukan prediksi SO_2 di Jakarta menggunakan *Artificial Neural Network* (*ANN*). Jenis algoritma ANN yang digunakan adalah *backpropagation*. Lebih lanjut, model prediksi dibangun dan dibandingkan berdasarkan tiga fungsi aktivasi dan skema pembagian data yang berbeda untuk memperoleh struktur atau arsitektur model ideal dengan tujuan mengoptimalkan hasil prediksi. Model dievaluasi menggunakan nilai *Mean Absolute Percentage Error* (MAPE) terkecil. Hasil penelitian menunjukkan bahwa model terbaik untuk prediksi indeks kandungan SO_2 adalah model yang menerapkan fungsi aktivasi Tanh dengan skema pembagian data 90% data pelatihan dan 10% data pengujian. Model tersebut memperoleh nilai MAPE sebesar 15.87412%, dan akurasi sebesar 84.12588%. Hal ini mengindikasikan bahwa model memiliki tingkat akurasi yang cukup baik dalam memprediksi indeks kandungan SO2 di Ibu Kota Jakarta dan dapat dijadikan sebagai pendekatan alternatif dalam meningkatkan efektivitas pengendalian SO_2.

Kata kunci: Indeks Kandungan SO_2, Prediksi, *Artificial Neural Network*, *Backpropagation*

Best Structural Analysis of Neural Network with Backpropagation Algorithm in Predicting Sulfide Content Index (SO2) in the Capital City of Jakarta

Abstract

Air pollution is an environmental problem commonly occurring in big cities in the country, including the capital city of Jakarta. The high population in Jakarta causes air pollution concentrations to increase due to the increased number of vehicles. This condition is exacerbated by the large amount of waste produced by factories from various industries. Sulfur Dioxide (SO_2) is one of the pollutants with the highest concentration in Jakarta, with a total emission load of 19.7 kton. Therefore, predicting the SO_2 index is an important research issue because SO_2 can impact various factors, such as the environment, agriculture and health. This research aims to find the best model for predicting SO_2 in Jakarta using an Artificial Neural Network (ANN). The type of ANN algorithm used is backpropagation. Furthermore, prediction models are built and compared based on three activation functions and data splitting schemes to obtain the ideal model structure or architecture to optimize prediction results. The model is evaluated using the smallest Mean Absolute Percentage Error (MAPE) value. The research results show that the best model for predicting the SO_2 content index is a model that applies the Tanh activation function with a data division scheme of 90% training data and 10% testing data. This model obtained a MAPE value of 15.87412% and an accuracy of 84.12588%. That indicates that the model has a pretty good level of accuracy in predicting the SO_2 index in the capital city of Jakarta and can be used as an alternative approach to increasing the effectiveness of SO_2 control.

Keywords: SO_2 Index, Prediction, Artificial Neural Network, Backpropagation

I. PENDAHULUAN

Kehidupan perkotaan, yang ditandai dengan padatnya lalu lintas kendaraan dan aktivitas industri, menghadirkan serangkaian permasalahan yang cukup kompleks. Pasalnya, udara merupakan salah satu faktor terpenting bagi kelangsungan hidup manusia [1]. Namun, mobilitas transportasi dan aktivitas industri yang tidak terkendali telah mengakibatkan produksi emisi polutan berlebihan sehingga kualitas udara menurun drastis [2]. Hal ini berpotensi mempengaruhi kesehatan dan keselamatan hidup manusia. Kualitas udara yang buruk juga dapat menyebabkan masalah lingkungan kontemporer lainnya seperti pemanasan global, hujan asam, berkurangnya jarak pandang, kabut asap, pembentukan aerosol, perubahan iklim, dan kematian dini [3].

Beberapa tahun terakhir, polusi udara merupakan masalah lingkungan yang umum terjadi di kota-kota besar di tanah air, tidak terkecuali Ibu Kota Indonesia, Jakarta. Daerah Khusus Ibukota Jakarta atau dikenal sebagai DKI Jakarta, merupakan sebuah provinsi yang secara geografis terbagi menjadi 5 kota dan 1 kabupaten yang meliputi Kepulauan Seribu. Berdasarkan sensus penduduk tahun 2020, jumlah penduduknya mencapai 10,7 juta jiwa dengan luas wilayah ±661,52 km^2. Tingginya jumlah penduduk menyebabkan peningkatan jumlah kendaraan sehingga mengakibatkan konsentrasi polusi udara semakin tinggi.

Kondisi ini diperburuk dengan kenyataan bahwa saat ini Jakarta menjadi episentrum segala politik, perekonomian, dan berbagai aspek vital Indonesia. Banyak orang yang bercita-cita mencari penghidupan dengan memanfaatkan keuntungan finansial yang ditawarkan oleh daerah perkotaan, sehingga mengakibatkan perpindahan penduduk pedesaan. Industri melihat hal ini sebagai peluang untuk mendirikan pabrik dan memperburuk polusi udara karena pengelolaan limbah yang tidak tepat [4]. Selain itu, terdapat pengaruh timbal balik antara pencemaran udara di kawasan Bogor, Depok, Bekasi, dan Tangerang (Bodetabek) yang merupakan daerah suburban DKI Jakarta dengan DKI Jakarta [5].

Upaya pemantauan kualitas udara di DKI Jakarta salah satunya dilakukan dengan cara mengukur kualitas udara menggunakan suatu indeks yang dikenal dengan nama Indeks Standar Pencemar Udara (ISPU). Pada tahun 2020, pemerintah Republik Indonesia menetapkan bahwa kualitas udara dihitung berdasarkan konsentrasi tujuh parameter (polutan udara) antara lain Nitrogen Dioksida (NO_2), Sulfida (SO_2), Karbon Monoksida (CO), Ozon (O_3), bahan partikulat berdiameter 10 mikron atau kurang (PM_{10}), partikel yang berdiameter 2,5 mikron atau lebih kecil ($PM_{2.5}$), dan Hidrokarbon (HC) [6]. Di antara banyaknya polutan udara tersebut, penelitian oleh Lestari dkk. [7] menunjukkan bahwa SO_2 merupakan salah satu polutan dengan konsentrasi tertinggi dengan total beban emisi sebesar 19.7 kton. Sebagian besar SO_2 dihasilkan dari pembakaran industri dan menyumbang hampir 67% dari total emisi.

Zat SO_2 dapat menimbulkan kerusakan pada berbagai material, benda maupun tanaman jika bereaksi dengan uap air di udara dan membentuk H_2SO_4 (hujan asam). Sulfida juga dapat menyebabkan iritasi saluran pernapasan, penurunan fungsi paru-paru, serta peningkatan gejala asma [8]. Oleh karena itu, prediksi tingkat pencemaran udara, khususnya parameter SO_2, merupakan isu penelitian yang penting karena berdampak terhadap berbagai faktor, seperti lingkungan, pertanian, dan kesehatan [9].

Prediksi adalah memperkirakan kemungkinan yang akan terjadi di masa depan berdasarkan data historis yang ada. Ada berbagai metode pemodelan yang umum digunakan untuk melakukan prediksi, salah satunya yaitu deret waktu. Namun, model deret waktu rentan terhadap *overfitting* dan jika *outlier* tidak ditangani dengan benar, hal ini dapat menyebabkan hasil prediksi yang diperoleh tidak akurat [10].

Dengan pesatnya kemajuan kecerdasan buatan dan pembelajaran mesin dalam beberapa tahun terakhir, model prediksi berbasis kecerdasan buatan semakin populer dan menarik lebih banyak perhatian [11]. Model prediksi yang umum digunakan adalah *Artificial Neural Network* (*ANN*), di mana tata nama dan bentuknya terinspirasi oleh jaringan syaraf di otak manusia [12].

Algoritma ANN yang umum digunakan adalah *backpropagation* [13, 14]. *Backpropagation* merupakan jaringan *multi-layer* yang memperbarui nilai bobot selama proses pelatihan dengan cara mundur dari lapisan *output* ke lapisan *input*. Dengan mekanisme tersebut, arsitektur *backpropagation* terbukti mampu menganalisis pola data historis secara lebih detail dan memperoleh hasil dengan kesalahan minimum [15].

Implementasi ANN mencakup spektrum yang luas dari berbagai bidang penelitian, termasuk untuk prediksi. Maleki dkk. [16] mengusulkan model ANN untuk memprediksi konsentrasi polutan udara per jam di Ahvaz, Iran, selama satu tahun penuh (Agustus 2009 - Agustus 2010) berdasarkan dua indeks kualitas udara, yaitu indeks pencemaran udara, dan indeks kesehatan kualitas udara. Hasil prediksi mereka menemukan bahwa model ANN dapat digunakan sebagai alternatif yang efektif untuk pengambilan keputusan yang lebih baik di bidang kualitas udara perkotaan.

Agarwal dkk. [17] mengembangkan ANN untuk prediksi kualitas udara jangka pendek di 32 lokasi berbeda di Delhi dengan memperkirakan konsentrasi polutan PM_{10}, $PM_{2.5}$, NO_2, dan O_3. Model ANN yang dibangun menunjukkan kinerja yang sangat baik dalam memprediksi semua polutan pada berbagai metrik evalusasi. Hamdan dkk. [18] membangun model ANN untuk memantau kualitas udara di Kementerian Lingkungan Hidup, Amman. Hasilnya menunjukkan bahwa pihak yang bertanggung jawab atas pengelolaan kualitas udara perkotaan dan pengambil keputusan dapat memanfaatkan ANN untuk memperkirakan konsentrasi polutan dan indeks kualitas udara.

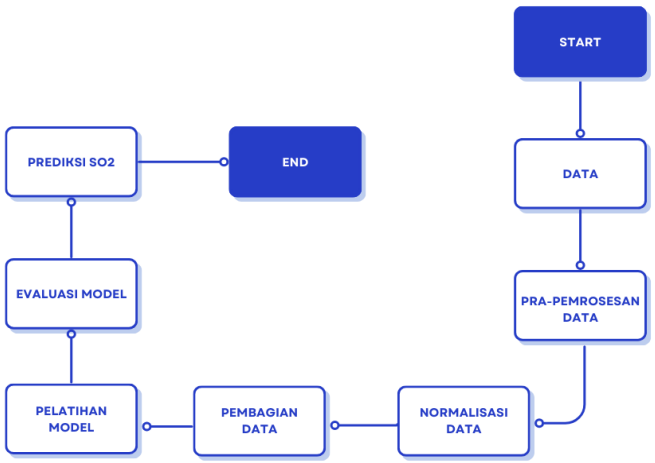
Vinas dkk. [19] memprediksi kualitas udara di Filipina menggunakan model ANN berbasis *feed forward*. Model prediksi tersebut menghasilkan kinerja terbaik dengan nilai MSE, MAE, MAPE yang minimum dan nilai R^2 sebesar 0.99. Sachdeva dkk. [20] mengusulkan kerangka kerja baru dan terintegrasi yang terdiri dari 4

modul berbeda untuk memprediksi indeks kualitas udara menggunakan data konsentrasi polutan dan data meteorologi. Metode yang digunakan untuk meramalkan polutan dan memprediksi kualitas udara pun beragam, di antaranya ANN, *Decision Tree Regression*, *Long-Short Term Memory*, *Random Forest Regression*, dan *k-Nearest Neighbour*. Namun, di antara metode lainnya, model ANN merupakan model menjanjikan yang mampu memprediksi semua polutan secara akurat.

Berdasarkan uraian di atas, salah satu polutan dengan konsentrasi tertinggi di Jakarta adalah SO_2. Kemudian, agar dapat meningkatkan efektivitas pengendalian SO_2 perlu dilakukan prediksi. Model ANN memiliki performa yang baik untuk prediksi indeks kualitas udara atau di Indonesia dikenal dengan nama Indeks Standar Pencemar Udara (ISPU). Oleh karena itu, penelitian ini bertujuan untuk memprediksi indeks polutan SO_2 menggunakan model ANN. Data yang diprediksi yaitu data SO_2 di Ibu Kota Jakarta. Model dibangun dengan algoritma *backpropagation* dan dibandingkan berdasarkan tiga fungsi aktivasi dan skema pembagian data yang berbeda untuk memperoleh struktur atau arsitektur model ideal dengan tujuan mengoptimalkan hasil prediksi. Lebih lanjut, model dievaluasi menggunakan nilai *Mean Absolute Percentage Error* (MAPE) terkecil.

II. METODOLOGI PENELITIAN

Pada bagian ini, tahapan penelitian diuraikan melalui diagram alir sebagai berikut:



Gambar 1. Diagram Alir Penelitian

Berdasarkan Gambar 1, tahapan penelitian terdiri dari pra-pemrosesan data, normalisasi, pembagian data, pelatihan model ANN, dan evaluasi model.

A. Data

Data yang digunakan adalah data sekunder yang diperoleh dari https://data.jakarta.go.id/dataset/ mengenai data histori indeks standar pencemaran udara di Provinsi DKI Jakarta selama 16 bulan yang terhitung sejak April 2020 sampai dengan Juli 2021. Data tersebut terdiri dari 6 parameter polutan udara, diantaranya: PM_{10}, PM_{25}, SO_2, NO_2, CO, dan O_3. Variabel-variabel yang ada kemudian diseleksi untuk menentukan variabel apa saja yang sesuai dengan penelitian ini. Adapun variabel yang sesuai, yaitu PM_{10}, SO_2, NO_2, CO, dan O_3.

B. Pra-pemrosesan Data

Pra-pemrosesan data mengacu pada tindakan pembersihan, modifikasi, dan transformasi data mentah sebelum diproses dan dianalisis. Tahapan ini merupakan langkah penting yang bertujuan untuk meningkatkan kualitas data dan mengurangi potensi bias [21].

Langkah awal pra-pemrosesan data yang dilakukan pada penelitian ini, adalah melakukan uji korelasi antar variabel yang ada guna mengidentifikasi variabel apa saja yang memiliki korelasi dengan polutan SO_2 serta menentukan *input* dan *output*.

Langkah kedua, dilakukan konversi data menjadi format deret waktu. Konversi data bertujuan untuk memanipulasi data agar sesuai dengan format data yang dibutuhkan model.

Langkah ketiga, data yang telah dikonversi menjadi format deret waktu dilakukan pengecekan *missing data* atau data yang hilang. Data yang hilang menghadirkan berbagai masalah. Pertama, hilangnya data mengurangi kekuatan statistik, yang menunjukkan probabilitas bahwa pengujian akan menolak hipotesis nol secara keliru. Kedua, data yang hilang dapat menimbulkan bias dalam estimasi parameter. Ketiga, dapat mengurangi keterwakilan sampel. Keempat, hal ini mungkin menimbulkan kompleksitas dalam analisis penelitian [22].

Oleh karena itu, sangat penting untuk menerapkan prosedur sistematis guna mengetahui keberadaan data yang hilang. Apabila terdapat *missing data* pada kumpulan data yang digunakan, maka langkah selanjutnya adalah memperbaiki data tersebut. Sebaliknya, jika tidak terdapat *missing data*, maka data akan melalui proses normalisasi pada tahapan berikutnya.

C. Normalisasi Data

Normalisasi data mengubah nilai fitur dalam kumpulan data awal menjadi rentang tertentu dengan tujuan mengurangi bias yang disebabkan oleh fitur-fitur yang kontribusi numeriknya lebih besar [23, 24]. Pada saat yang sama, normalisasi data dapat mempercepat proses pembelajaran karena fitur-fitur pada data diukur pada skala yang sama [25].

Meskipun terdapat berbagai teknik normalisasi data, beberapa penelitian menunjukkan bahwa *min-max normalization* adalah metode terbaik untuk menormalisasi data [26–28].

Min-max normalization mengubah nilai fitur pada data menjadi rentang antara 0 dan 1 melalui Persamaan (1) sebagai berikut:

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

dengan x_i adalah data awal, x_{min} adalah nilai minimum, x_{max} adalah nilai maksimum, dan x' adalah data yang telah dinormalisasi.

D. Pembagian Data

Pada saat membangun model pembelajaran, kumpulan data biasanya dibagi menjadi data pelatihan dan data pengujian. Data pelatihan digunakan untuk menyesuaikan dan membaca pola data yang tidak diketahui sebelumnya. Model pembelajaran kemudian dievaluasi menggunakan data pengujian untuk memastikan keakuratan hasil prediksi [29].

Rasio pembagian data yang umum digunakan adalah 80% data pelatihan dan 20% data pengujian. Tetapi pada praktiknya, rasio lain seperti 90% : 10%, 70% : 30%, dan 60% : 40% juga banyak digunakan karena tidak ada aturan baku mengenai rasio terbaik atau optimal untuk kumpulan data tertentu [30]. Oleh sebab itu, untuk mengetahui rasio optimal untuk prediksi SO_2 di Jakarta, penelitian ini akan membagi data menjadi empat skema pembagian data, yaitu 90% : 10%, 80% : 20%, 70% : 30%, dan 60% : 40%.

E. *Artificial Neural Network (ANN)*

Artificial Neural Network, atau sering disebut *Neural Network*, merupakan model tingkat lanjut hasil integrasi sejumlah besar ekspresi matematika yang relatif sederhana (fungsi aktivasi) dengan variabel masukan untuk memberikan prediksi keluaran berdasarkan parameter dan arsitektur jaringan [31]. Berbagai permasalahan yang kompleks dalam berbagai aplikasi seperti optimasi, simulasi, pemodelan, *clustering*, pengenalan pola, klasifikasi, dan prediksi terbukti mampu diatasi dengan baik oleh model ANN [32].

Sesuai dengan namanya, arsitektur ANN terdiri dari *neuron-neuron* yang saling berhubungan. Setiap *neuron* memiliki fungsi aktivasi yang berfungsi untuk menentukan nilai *output* dari *neuron*. Fungsi aktivasi yang akan digunakan ada tiga, yaitu: Sigmoid, Tanh, dan ReLU. Putra dkk. [33] mendeskripsikan ketiga fungsi aktivasi sebagai berikut:

1. Fungsi Sigmoid: atau dikenal sebagai fungsi logistik. Fungsi ini memetakan nilai input ke dalam rentang 0 hingga 1. Persamaan fungsi aktivasi dirumuskan pada Persamaan (2).

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

2. Fungsi Tanh: fungsi aktivasi alternatif untuk Sigmoidkarena karakteristik keduanya mirip. Namun, fungsi ini memiliki rentang nilai yang lebih luas, yaitu antara -1 dan 1. Oleh karena itu, fungsi Tanh cocok untuk pemodelan *nonlinear* yang kompleks. Formula matematis dari fungsi Tanh diuraikan pada Persamaan (3) berikut:

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \quad (3)$$

3. Fungsi *Rectified Linear Unit* (ReLU): fungsi aktivasi yang digunakan untuk menormalkan nilai yang dihasilkan *neuron*. Persamaan fungsi ReLU ditunjukkan pada Persamaan (4).

$$f(x) = \max(0, x) \quad (4)$$

Berdasar Persamaan (4), diketahui bahwa fungsi ReLU akan mengubah nilai *x* menjadi 0 jika $x \leq 0$. Sedangkan, jika nilai $x > 0$, maka $x = x$.

F. Evaluasi Model

Apabila proses pelatihan telah selesai, tahapan selanjutnya adalah evaluasi model. Evaluasi model bertujuan untuk menilai performa model prediksi. Metrik yang paling tepat untuk tujuan ini dibanding metrik evaluasi lainnya adalah MAPE [34].

Metrik ini mengukur performa model prediksi dengan cara membagi setiap *error* dalam model dengan nilai aktual dan menampilkan hasilnya sebagai rata-rata dalam persentase. Secara matematis, MAPE didefinisikan pada Persamaan (5) sebagai berikut:

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

Berdasarkan nilai MAPE, tingkat keakuratan model diklasifikasikan menjadi empat kategori dan dideskripsikan pada Tabel 1 sebagai berikut [35]:

TABEL I. TINGKAT KEAKURATAN MODEL BERDASARKAN NILAI MAPE

MAPE (%)	Kemampuan Prediksi
<10	Sangat Akurat
10-20	Akurat
20-50	Cukup Akurat
>50	Tidak Akurat

G. Prediksi

Tahapan terakhir dari penelitian ini adalah melakukan prediksi SO_2 berdasarkan model terbaik yang diperoleh selama proses pelatihan. Namun sebelum itu, keluaran data hasil prediksi harus didenormalisasi.

Denormalisasi merupakan kebalikan dari proses normalisasi. Denormalisasi digunakan sebagai prosedur untuk mengembalikan nilai data yang sebelumnya telah melalui proses normalisasi ke nilai data sebenarnya.

Terakhir, data hasil prediksi akan dibandingkan dengan data aktual untuk memastikan hasil prediksi menggunakan ANN sesuai dengan data sebenarnya.

III. HASIL DAN DISKUSI

A. Pra-pemrosesan Data

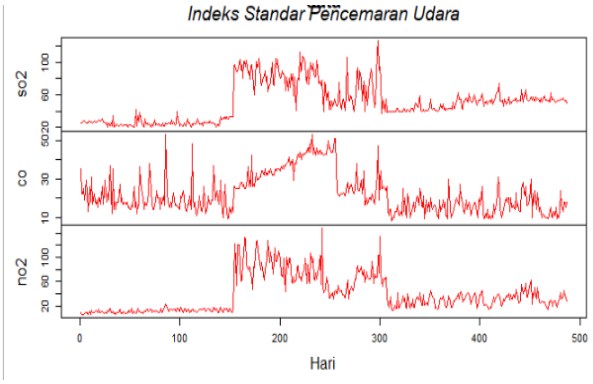
Tahapan awal pada penelitian ini adalah pra-pemrosesan data. Namun, sebelumnya perlu dilakukan identifikasi korelasi antar variabel untuk menentukan variabel mana yang akan digunakan sebagai *input* dan *output* dengan melakukan uji korelasi terhadap variabel-variabel yang ada.

```
> cor(data)
             so2
so2   1.0000000
co    0.5316930
no2   0.8665955
pm10  0.1069343
o3    0.0258657
```

Gambar 2. Nilai Koefisien Korelasi Antar Variabel

Berdasarkan Gambar 2, variabel yang menunjukkan korelasi kuat dengan variabel SO_2 adalah variabel CO dan NO_2. Oleh karena itu, kedua indeks ini akan digunakan sebagai variabel *input*. Sedangkan indeks kandungan SO_2 akan digunakan sebagai variabel *output* karena tujuan penelitian ini adalah untuk memprediksi data indeks kandungan SO_2.

Data tersebut kemudian dikonversi ke dalam format deret waktu dan divisualisasikan dalam bentuk plot, seperti pada Gambar 3 berikut:



Gambar 3. Visualisasi Data ISPU

Gambar 3 menginformasikan bahwa indeks kandungan SO_2, CO, dan NO_2 menunjukkan pergerakan yang fluktuatif. Hal ini mengindikasikan bahwa data cenderung naik dan turun setiap harinya. Oleh karena itu, penting untuk melakukan prediksi indeks kandungan SO_2 guna meningkatkan efektivitas pengendalian polutan.

Langkah selanjutnya adalah memastikan tidak adanya informasi yang hilang pada data melalui algoritma berikut:

```
> missing(data)
[1] FALSE
```

Gambar 4. Cek *Missing Value*

Gambar 4 menunjukkan bahwa tidak ada informasi yang hilang pada data yang digunakan. Sehingga, data tidak perlu dimodifikasi dan dapat dilanjutkan ke tahapan berikutnya yaitu, normalisasi data.

B. Normalisasi Data

Normalisasi data merupakan salah satu bentuk proses transformasi data yang bertujuan untuk menormalisasi nilai numerik suatu kumpulan data dan memastikan tidak ada variabel data yang mendominasi karena perbedaan rentang nilai. Penelitian ini menggunakan teknik normalisasi *min-max normalization* yang hasilnya diilustrasikan pada Gambar 5 berikut:

```
> summary(normalx)
      so2               co               no2        
 Min.   :0.00000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.07547   1st Qu.:0.1556   1st Qu.:0.05634  
 Median :0.27358   Median :0.2444   Median :0.16197  
 Mean   :0.28585   Mean   :0.3136   Mean   :0.22836  
 3rd Qu.:0.37264   3rd Qu.:0.4000   3rd Qu.:0.34155  
 Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
```

Gambar 5. Hasil *Min-Max Normalization*

C. Pembagian Data

Setelah rentang nilainya sama, data dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian. Pada penelitian ini, data dibagi menjadi empat skema pembagian data yang berbeda dengan tujuan memperoleh hasil terbaik berdasarkan teknik membagi data. Skema pembagian yang digunakan disajikan pada Tabel 2 berikut:

TABEL II. PEMBAGIAN DATA PELATIHAN DAN DATA PENGUJIAN

Pembagian Data	Data pelatihan	Data pengujian
60% dan 40%	292	195
70% dan 30%	341	146
80% dan 20%	390	97
90% dan 10%	438	49

D. Pelatihan Model

Salah satu kendala yang dihadapi dalam mencapai keberhasilan penerapan ANN adalah penentuan struktur jaringan, yang erat kaitannya dengan jumlah *nodes* atau *neuron* pada *hidden layer* dan jenis fungsi aktivasi yang digunakan. Heaton (2017) menyatakan bahwa jumlah nodes ideal pada masing-masing *hidden layer* dapat ditentukan dengan aturan sebagai berikut:

1. Jumlah *nodes* pada *hidden layer* kurang dari dua kali lipat jumlah *input layer*.
2. Jumlah *nodes* pada *hidden layer* bernilai $\frac{2}{3}$ dari jumlah *input layer* ditambah dengan *output layer* $\left(\frac{2}{3}(input + output)\right)$.
3. Jumlah *nodes* pada *hidden layer* berada di antara ukuran *input layer* dan *output layer*.

Berdasarkan aturan di atas, jumlah *nodes* pada masing-masing *hidden layer* yang memenuhi kriteria yaitu tiga, dua, dan satu. Hal ini dikarenakan *input layer* pada model yang dibangun memiliki dua *nodes*. Sehingga, jumlah ideal dari *hidden layer* untuk penelitian ini adalah tiga, dengan

masing-masing *hidden layer* terdiri dari tiga *nodes*, dua *nodes* dan satu *nodes*.

Terakhir, *neuron* pada setiap *hidden layer* diaktifkan dengan tiga fungsi aktivasi, yaitu Sigmoid, Tanh, dan ReLU.

E. Evaluasi Model

Hasil pelatihan kemudian dibandingkan berdasarkan nilai MAPE dan akurasi yang diperoleh selama proses pelatihan dan ditampilkan pada Tabel 3 sebagai berikut:

TABEL III. AKURASI MODEL

Pembagian Data pelatihan dan Uji	Fungsi Aktivasi	*MAPE*	Akurasi
60% dan 40%	Sigmoid	54.8380	45.1619
	Tanh	22.8907	77.1092
	ReLU	50.8370	49.1629
70% dan 30%	Sigmoid	42.8029	57.1970
	Tanh	22.2673	77.7326
	ReLU	50.8370	49.1629
80% dan 20%	Sigmoid	57.4757	42.5242
	Tanh	17.7313	82.2686
	ReLU	50.8370	49.1629
90% dan 10%	Sigmoid	51.7639	48.2360
	Tanh	15.8741	84.1258
	ReLU	50.8370	49.1629

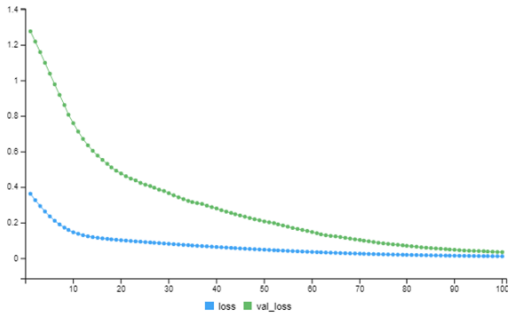
Berdasarkan Tabel 3, dapat diketahui bahwa model yang menerapkan Tanh sebagai fungsi aktivasi, memiliki tingkat akurasi prediksi yang lebih tinggi dibandingkan model lain yang menerapkan fungsi aktivasi Sigmoid atau ReLU meskipun menggunakan skema pembagian data yang sama. Misalnya, dengan skema pembagian data 60% data pelatihan dan 40% data pengujian, model dengan fungsi Sigmoid dan fungsi ReLU menghasilkan nilai MAPE masing-masing sebesar 54.8380% dan 50.8370%. Nilai MAPE yang melebihi 50% tersebut menunjukkan bahwa prediksi model tidak akurat.

Sementara pada fungsi Tanh, meskipun skema pembagian data yang digunakan tetap sama, yaitu 60% data pelatihan dan 40% data pengujian, nilai MAPE yang diperoleh lebih kecil, yaitu 22.8907%. Hal ini mengindikasikan bahwa model yang memanfaatkan Tanh sebagai fungsi aktivasi memiliki tingkat akurasi yang cukup baik untuk prediksi indeks kandungan SO_2.

Selain dipengaruhi oleh fungsi aktivasi yang digunakan, Tabel 3 juga memberikan informasi bahwa tingkat keakuratan model dapat dipengaruhi oleh teknik pembagian data. Semakin besar rasio data yang digunakan untuk melatih model, maka semakin kecil nilai MAPE yang dihasilkan. Oleh karena itu, berdasarkan Tabel 3 dapat disimpulkan bahwa, model terbaik untuk prediksi indeks kandungan SO_2 adalah model yang menggunakan fungsi aktivasi Tanh dengan 90% data pelatihan dan 10% data pengujian.

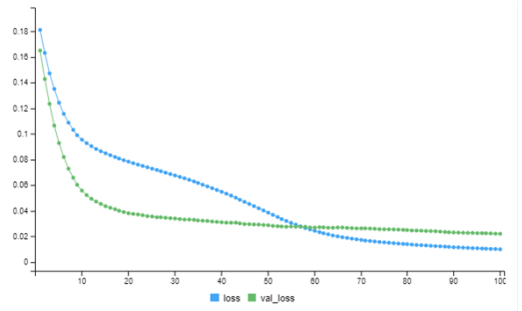
Meskipun demikian, untuk memastikan bahwa model tidak hanya akurat untuk data pelatihan tetapi juga memiliki generalisasi yang baik untuk data baru, diperlukan pendekatan lain. Pendekatan lain yang dapat digunakan, yaitu dengan mengamati grafik *loss* suatu model.

Grafik *loss* yang akan diamati adalah grafik *loss* dari model yang menerapkan Tanh sebagai fungsi aktivasi. Model lain yang diusulkan tidak perlu diamati grafik *loss* nya, dikarenakan model-model tersebut tidak memenuhi kriteria untuk menjadi prediktor yang baik. Grafik *loss* untuk masing-masing model yang menerapkan fungsi aktivasi Tanh divisualisasikan melalui Gambar 6, Gambar 7, Gambar 8, dan Gambar 9.



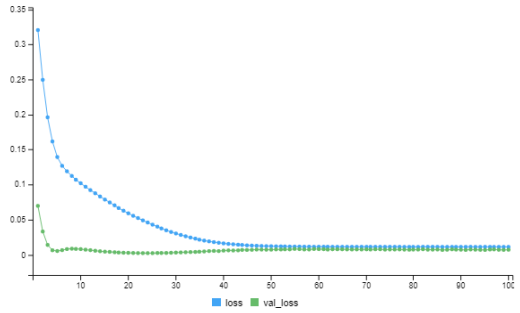
Gambar 6. Grafik *Loss* Fungsi Tanh dengan 60% Data pelatihan dan 40% Data pengujian

Gambar 6 menunjukkan bahwa model yang dibangun dengan skema 60% data pelatihan dan 40% data pengujian mengalami *underfitting*. *Underfitting* merupakan suatu kondisi di mana model memiliki performa yang buruk selama proses pelatihan maupun pengujian. Hal ini dikarenakan nilai *loss* dan *validation loss* yang diperoleh belum konvergen pada titik tertentu.

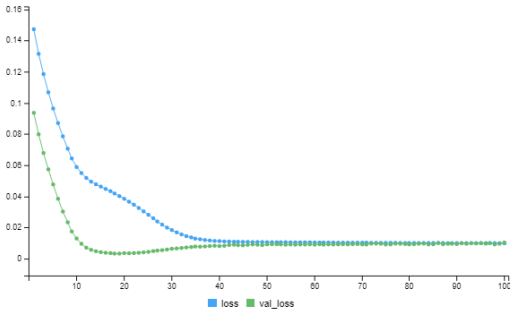


Gambar 7. Grafik *Loss* Fungsi Aktivasi Tanh dengan 70% Data pelatihan dan 30% Data pengujian

Berdasarkan Gambar 7, diketahui bahwa nilai *loss* dan *validation loss* dari model yang menggunakan skema 70% data pelatihan dan 30% data pengujian telah konvergen pada titik tertentu. Namun, kemudian rentang nilai *loss* dan *validation loss* tersebut cenderung menjauh. Kondisi ini disebut *overfitting*. Artinya, walaupun model tersebut menunjukkan performa yang baik selama pelatihan, namun model tidak cukup baik untuk digunakan sebagai memprediksi data baru.



Gambar 8. Grafik *Loss* Fungsi Aktivasi Tanh dengan 80% Data pelatihan dan 20% Data pengujian



Gambar 9. Grafik *Loss* Fungsi Aktivasi Tanh dengan 90% Data pelatihan dan 10% Data pengujian

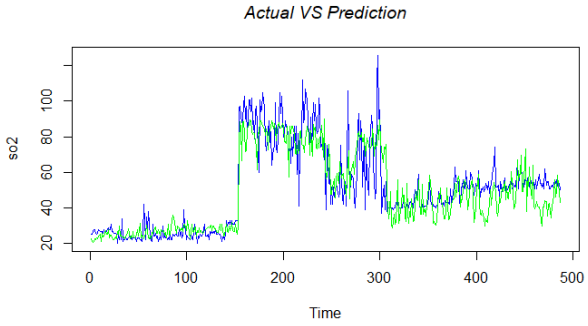
Berdasarkan analisis dari Gambar 8 dan Gambar 9, diketahui bahwa kedua model ini merupakan model dengan grafik *loss* yang fit. Grafik *loss* yang diperoleh telah konvergen, menunjukkan bahwa model tidak mengalami *underfitting* maupun *overfitting*. Hal ini menyiratkan bahwa performa model sangat baik selama pelatihan. Di samping itu, model juga memiliki kemampuan generalisasi yang baik untuk data baru.

Akan tetapi, dari kedua grafik *loss* yang di amati, model yang paling ideal dari keduanya adalah model yang menerapkan 90% data pelatihan dan 10% data pengujian. Hal ini dikarenakan, nilai *loss* dan validation *loss* dari model tersebut cenderung sama, sebagaimana ditunjukkan oleh garis biru dan garis hijau yang sangat berhimpitan. Selain itu, nilai MAPE dari model ini juga lebih kecil, yaitu 15.8741%. Di sisi lain, model dengan skema pembagian 80% data pelatihan dan 20% data pengujian, memperoleh nilai MAPE sebesar 17.7313%.

Dengan demikian, berdasarkan tingkat keakuratan model dan grafik *loss* yang telah diamati, dapat disimpulkan bahwa model terbaik untuk prediksi indeks kandungan SO_2 pada ISPU adalah model yang menerapkan fungsi aktivasi Tanh dengan skema pembagian data 90% data pelatihan dan 10% data pengujian. Hasil ini menunjukkan bahwa rasio data, dan pemilihan fungsi aktivasi yang digunakan dapat mempengaruhi keakuratan dan kinerja model.

F. Hasil Prediksi SO_2

Setelah model terbaik telah diperoleh, maka langkah selanjutnya adalah melakukan prediksi indeks kandungan SO_2 pada ISPU untuk memperkirakan kualitas udara di Ibu Kota Jakarta. Gambar 10 merupakan ilustrasi perbandingan hasil prediksi dengan data sebenarnya menggunakan model terbaik yang diusulkan. Data sebenarnya direpresentasikan oleh garis berwarna biru. Sedangkan, data hasil prediksi direpresentasikan oleh garis berwarna hijau.



Gambar 10. Plot *Actual VS Prediction*

Berdasarkan Gambar 10, dapat dikatakan bahwa pola data prediksi mengikuti pola sebaran data sebenarnya. Meskipun masih terdapat perbedaan yang cukup signifikan, model ANN yang diusulkan terbukti mampu melakukan prediksi indeks kandungan SO2 dengan baik. Hasil prediksi dan nilai-nilainya yang telah didenormalisasi untuk rentang waktu April 2020 - Juli 2021 dipaparkan pada Tabel 4 berikut:

TABEL IV. DATA AKTUAL DAN HASIL PREDIKSI SO_2

Tanggal	Aktual SO_2	Prediksi SO_2
4/01/2020	25	22.86283
4/02/2020	25	21.33347
4/03/2020	27	20.46011
4/04/2020	28	20.95354
4/05/2020	28	22.65526
4/06/2020	26	22.34667
…	…	…
…	…	…
7/26/2021	54	46.24096
7/27/2021	56	49.2434
7/28/2021	53	58.55529
7/29/2021	52	53.01764
7/30/2021	54	49.1945
7/31/2021	50	43.15153

IV. KESIMPULAN DAN SARAN

Model terbaik untuk prediksi indeks kandungan SO_2 adalah model yang menerapkan Tanh sebagai fungsi aktivasi. Keakuratan model tersebut juga dipengaruhi oleh teknik pembagian data yang digunakan, yaitu 90% data pelatihan dan 10% data pengujian. Di samping itu, arsitektur ideal untuk model yang diusulkan mencakup dua *nodes* input layer, tiga *hidden layer* dengan tiga *nodes* pada

hidden layer pertama, dua *nodes* pada *hidden layer* kedua dan satu *nodes* pada *hidden layer* ketiga serta *output layer*.

Selama proses evaluasi, model tersebut memperoleh nilai MAPE sebesar 15.87412%, dan akurasi sebesar 84.12588%. Hasil ini menunjukkan bahwa model memiliki tingkat akurasi yang cukup baik dalam memprediksi indeks kandungan SO2 di Ibu Kota Jakarta.

Meskipun demikian, model ini berpotensi untuk dilakukan pengembangan lebih lanjut guna memperoleh performa model dan tingkat keakuratan yang lebih baik. Sebagai contoh, untuk pengembangan kedepannya akan dilakukan pembaruan dalam membagi data dengan menggunakan *k-fold Cross Validation*. Melalui pendekatan ini, setiap data memiliki kesempatan menjadi data pelatihan sekaligus data pengujian, sehingga dapat meningkatkan akurasi model dan memperoleh hasil prediksi yang lebih baik.

DAFTAR PUSTAKA

[1] S. Aswatha *et al.*, "Smart air pollution monitoring system," *Glob. Nest J.*, vol. 25, no. 3, pp. 125–129, 2023, doi: 10.30955/gnj.004396.

[2] M. A. Fath, "Pengaruh Kualitas Udara dan Kondisi Iklim terhadap Perekonomian Masyarakat (Literature Review)," *Media Gizi Kesmas*, vol. 10, no. 2, p. 329, 2021, doi: 10.20473/mgk.v10i2.2021.329-342.

[3] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *Int. J. Environ. Sci. Technol.*, vol. 20, no. 5, pp. 5333–5348, 2023, doi: 10.1007/s13762-022-04241-5.

[4] A. N. Anggraini, N. K. Ummah, Y. Fatmasari, and K. F. Hayati Holle, "Air Quality Forecasting in DKI Jakarta Using Artificial Neural Network," *Matics*, vol. 14, no. 1, pp. 1–5, 2022, doi: 10.18860/mat.v14i1.13863.

[5] N. S. Darmanto and A. Sofyan, "Analisis Distribusi Pencemar Udara No2, So2, Co, Dan O2 Di Jakarta Dengan Wrf-Chem," *J. Teh. Lingkung.*, vol. 18, no. 1, pp. 54–64, Apr. 2012, doi: 10.5614/jtl.2012.18.1.6.

[6] T. Handhayani, "An integrated analysis of air pollution and meteorological conditions in Jakarta," *Sci. Rep.*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-32817-9.

[7] P. Lestari, S. Damayanti, and M. K. Arrohman, "Emission Inventory of Pollutants (CO, SO2, PM2.5, and NOX) in Jakarta Indonesia," in *IOP Conference Series: Earth and Environmental Science*, 2020. doi: 10.1088/1755-1315/489/1/012014.

[8] Masito A, "Analisis Risiko Kualitas Udara Ambien (No2 Dan So2) Dan Gangguan Pernapasan Pada Masyarakat Di Wilayah Kalianak Surabaya," *J. Kesehat. Lingkung.*, vol. 10, no. 4, pp. 394–401, 2018.

[9] N. Eliyati, M. Rahmayani, S. Wijaya, D. A. Zayanti, E. S. Kresnawati, and Y. Resti, "Prediction of Air Quality Index Using Decision Tree With Discretization," *Indones. J. Eng. Sci.*, vol. 3, no. 3, pp. 061–067, 2022, doi: 10.51630/ijes.v3i3.82.

[10] A. S. Bharatpur, "A Literature Review on Time Series Forecasting Methods.," 2022.

[11] H. Liu, G. Yan, Z. Duan, and C. Chen, "Intelligent modeling strategies for forecasting air quality time series: A review," *Applied Soft Computing*, vol. 102. 2021. doi: 10.1016/j.asoc.2020.106957.

[12] R. Qamar and B. A. Zardari, "Artificial Neural Networks - An overview," *Mesopotamian Journal of Computer Science*, vol. 2023. pp. 130–139, 2023. doi: 10.58496/mjcsc/2023/015.

[13] J. T. Hardinata, M. Zarlis, E. B. Nababan, D. Hartama, and R. W. Sembiring, "Modification of Learning Rate with Lvq Model Improvement in Learning Backpropagation," in *Journal of Physics: Conference Series*, 2017. doi: 10.1088/1742-6596/930/1/012025.

[14] A. W. Putri, "Implementasi Artificial Neural Network (ANN) Backpropagation Untuk Klasifikasi Jenis Penyakit Pada Daun Tanaman Tomat," *MATHunesa J. Ilm. Mat.*, vol. 9, no. 2, pp. 344–350, 2021, doi: 10.26740/mathunesa.v9n2.p344-350.

[15] W. Setiawan, A. Barokah, and Mula'ab, "Rainfall Prediction Using Backpropagation with Parameter Tuning," *MATEC Web Conf.*, vol. 372, p. 07003, 2022, doi: 10.1051/matecconf/202237207003.

[16] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artificial neural network model," *Clean Technol. Environ. Policy*, vol. 21, no. 6, pp. 1341–1352, 2019, doi: 10.1007/s10098-019-01709-w.

[17] S. Agarwal *et al.*, "Air quality forecasting using artificial neural networks with real-time dynamic error correction in highly polluted regions," *Sci. Total Environ.*, vol. 735, 2020, doi: 10.1016/j.scitotenv.2020.139454.

[18] M. A. Hamdan, M. F. B. Ata, and A. H. Sakhrieh, "Air Quality Assessment and Forecasting Using Neural Network Model," *J. Ecol. Eng.*, vol. 22, no. 6, pp. 1–11, 2021, doi: 10.12911/22998993/137444.

[19] M. Joy, D. Viñas, B. D. Gerardo, and R. P. Medina, "Forecasting PM 2.5 and PM 10 Air Quality Index using Artificial Neural Network," 2022. [Online]. Available: http://journalppw.com

[20] S. Sachdeva, H. Singh, S. Bhatia, and P. Goswami, "An integrated framework for predicting air quality index using pollutant concentration and meteorological data," *Multimed. Tools Appl.*, 2023, doi: 10.1007/s11042-023-17432-0.

[21] G. Grolemund, *Data Wrangling with R*, no. January 2015. 2016. [Online]. Available: https://www.oreilly.com/library/

[22] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*,

vol. 64, no. 5. pp. 402–406, 2013. doi: 10.4097/kjae.2013.64.5.402.

[23] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, 2020, doi: 10.1016/j.asoc.2019.105524.

[24] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, vol. 10, no. 11, 2022, doi: 10.3390/math10111942.

[25] G. Aksu, C. O. Guzeller, and M. T. Eser, "The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model," *Int. J. Assess. Tools Educ.*, vol. 6, no. 2, 2019, doi: 10.21449/ijate.479404.

[26] A. Ambarwari, Q. J. Adrian, and Y. Herdiyeni, "Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, 2020, [Online]. Available: http://jurnal.iaii.or.id

[27] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[28] S. Sinsomboonthong, "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification," *Int. J. Math. Math. Sci.*, vol. 2022, 2022, doi: 10.1155/2022/3584406.

[29] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, 2022, doi: 10.1080/00401706.2021.1921037.

[30] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.

[31] H. Sildir, S. Sarrafi, and E. Aydin, "Optimal artificial neural network architecture design for modeling an industrial ethylene oxide plant," *Comput. Chem. Eng.*, vol. 163, 2022, doi: 10.1016/j.compchemeng.2022.107850.

[32] N. Alifiah, D. Kurniasari, Amanto, and Warsono, "Prediction of COVID-19 Using the Artificial Neural Network (ANN) with K-Fold Cross-Validation," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 9, no. 1, pp. 16–27, 2023, doi: 10.20473/jisebi.9.1.16-27.

[33] D. S. Putra, M. Azmi, Muslikhin, and W. Purwanto, "ANN Activation Function Comparative Study for Sinusoidal Data," in *Journal of Physics: Conference Series*, 2022. doi: 10.1088/1742-6596/2406/1/012029.

[34] M. Meng and C. Song, "Daily photovoltaic power generation forecasting model based on random forest algorithm for north China in winter," *Sustain.*, vol. 12, no. 6, 2020, doi: 10.3390/su12062247.

[35] E. Vivas, H. Allende-Cid, and R. Salas, "A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score," *Entropy*, vol. 22, no. 12. pp. 1–24, 2020. doi: 10.3390/e22121412.