Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

TUGAS AKHIR

Jodikal Pomalingo
41518010033

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Tugas Akhir

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh:
Jodikal Pomalingo
41518010033

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

LEMBAR PERNYATAAN ORISINALITAS

LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM	:	41518010033
Nama	:	Jodikal Pomalingo
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan didalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 23 Februari 2022

10000 METERAI TEMPEL CBF0AJX788185931

Jodikal Pomalingo

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Jodikal Pomalingo
NIM	:	41518010033
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul diatas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Mercu Buana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 23 Februari 2022

METERAI TEMPEL
3D115AJX788185932

Jodikal Pomalingo

SURAT PERNYATAAN LUARAN TUGAS AKHIR

SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Jodikal Pomalingo
NIM	:	41518010033
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis		Status	
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi		Diajukan	√
		Jurnal Nasional Terakreditasi	√		
		Jurnal International Tidak Bereputasi		Diterima	
		Jurnal International Bereputasi			
	Disubmit/dipublikasikan di :	Nama Jurnal	:Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)		
		ISSN	:2338-3070, e-ISSN :2338-3062		
		Link Jurnal	:http://journal.uad.ac.id/		
		Link File Jurnal Jika Sudah di Publish	:		

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 23 Februari 2022

METERAI TEMPEL 80D1AJX788185935

Jodikal Pomalingo

LEMBAR PERSETUJUAN PENGUJI

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010033
Nama	:	Jodikal Pomalingo
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 26 Maret 2022

(Ir. Emil R. Kaburuan, Ph. D., IPM.)

LEMBAR PERSETUJUAN PENGUJI

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010033
Nama	:	Jodikal Pomalingo
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 30 Maret 2022

(Umniy Salamah, ST, MMSI)

LEMBAR PERSETUJUAN PENGUJI

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010033
Nama	:	Jodikal Pomalingo
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 08 Maret 2022

(Vina Ayumi, S.Kom., M.Kom)

LEMBAR PENGESAHAN

LEMBAR PENGESAHAN

NIM	:	41518010033
Nama	:	Jodikal Pomalingo
Judul Tugas Akhir	:	Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 23 Februari 2022

Menyetujui,

(Dr. Devi Fitrianah, S.Kom., MTI.)
Dosen Pembimbing

Mengetahui,

(Wawan Gunawan, S.Kom, MT)	(Ir. Emil R. Kaburuan, Ph. D., IPM.)
Koord. Tugas Akhir Teknik Informatika	Ka. Prodi Teknik Informatika

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah SWT yang telah memberikan rahmat dan hidayah-Nya sehingga penulis bisa menyelesaikan tugas akhir yang berjudul "*Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study of Air Pollution Standard Index Data in DKI Jakarta*", sebagai persyaratan untuk menyelesaikan Program Sarjana Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Penulis menyadari tanpa adanya dukungan, bimbingan, dan bantuan dari pihak – pihak terkait, penulis tidak mungkin menyelesaikan tugas akhir ini. Untuk itu, penulis ingin mengucapkan terima kasih kepada:

1. Orang tua yang selalu berdoa dan memberikan dukungan kepada penulis sehingga dapat menyelesaikan tugas akhir dengan lancar.
2. Ibu Dr. Devi Fitrianah, M.TI sebagai Dosen Pembimbing Tugas Akhir untuk segala arahan serta saran sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik
3. Bapak Dr. Harwikarya, MT. sebagai Dosen Pembimbing Akademik yang membantu dalam mengikuti dan menyelesaikan studi penulis di Fakultas Ilmu Komputer
4. Seluruh staf pengajar di Universitas Mercu Buana khususnya Program Studi Teknik Informatika Fakultas Ilmu Komputer yang menjadi tempat penulis dalam menimbah ilmu selama menempuh Pendidikan di Universitas Mercu Buana.

Akhir kata, penulis sangat berharap penelitian ini bisa bermanfaat untuk para pembaca. Penulis menyadari bahwasannya dalam penelitian ini masih terdapat banyak kekurangan, oleh karena itu penulis membutuhkan saran dan masukkan yang konstruktif agar dapat menyempurnakan penelitian ini.

Jakarta, 05 Maret 2022

Jodikal Pomalingo

DAFTAR ISI

NASKAH JURNAL

Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study Of Air Pollution Standard Index Data in DKI Jakarta

Jodikal Pomalingo[1], Devi Fitrianah[2]
[1,2]Department of Informatics, Faculty of Computer Science Universitas Mercu Buana, Jakarta, Indonesia

ARTICLE INFO

Article history:

Received July x, 20xx
Revised August x, 20xx
Accepted September x, 20xx

Keywords:

Deep Learning;
LSTM;
1D CNN;
classification;
ISPU;

ABSTRACT

Many big cities such as DKI Jakarta are experiencing a clean air crisis due to the development of industrialization, the increase in the number of private cars, and the burning of fossil fuels; Air quality is declining, with air pollution getting worse. The problem of handling air pollution is still the government's main focus because the impact of air pollution affects daily life, causing severe health problems for humans and other living things. We need a system that can classify air quality based on air pollution standard index parameters to become a consideration for the government in making air pollution control decisions. Therefore, this study aims to classify air quality based on parameters that affect air quality based on air quality categories. In addition, it also compares the Long Short Term Memory (LSTM) and One Dimensional Convolutional Neural Network (1D CNN) algorithms. Based on the experiments that have been carried out, the LSTM algorithm outperforms CNN 1D. The results show that both algorithms provide significant accuracy results equally well. The cross-validation results show that the LSTM algorithm obtains the best accuracy of 98.67% on a 5-fold cross-validation with an execution time of 315,049s. At the same time, the CNN 1D algorithm obtained the best accuracy of 98.08% with a time of 416.74s. LSTM provides better accuracy values for fewer k-folds. In comparison, CNN 1D obtained a better accuracy value at a larger k-fold. In conclusion, with the type of quantitative data and the characteristics of the low level of data variation, LSTM can be the proper method in classifying air quality.

Jodikal Pomalingo,
Faculty of Computer Science Universitas Mercu Buana, Jl. Raya Meruya Selatan no. 1, Kembangan Jakarta Barat 16550, Jakarta, Indonesia
Email: 41518010033@student.mercubuana.ac.id

1. INTRODUCTION

Air is a basic need for the survival and development of all life on Earth. Therefore, air affects health and affects economic development. Many big cities like DKI Jakarta are currently suffering from smog, which has affected daily life and caused serious health problems. Based on IQAir Visual data, Jakarta is ranked in the top five cities in the world with the worst air quality[1]. As a result of the development of industrialization, the increase in the number of private cars, and the burning of fossil fuels, decreased air quality, with increasingly severe air pollution. Air pollution consists of a mixture of gases and particulates in harmful quantities released into the atmosphere due to natural or human activities[2], [3]. A system is needed to predict future air quality automatically based on Air Pollution Standard Index (ISPU) data. It is hoped that the air quality prediction system can provide accurate information about the future pollution situation, which is effectively useful for efficient air pollution control operations and helps plan prevention [3], [4].

Until now, various studies in the field of environment and data have proposed air quality prediction methods, especially data classified into statistical methods, deep learning methods[5], and machine learning methods[2]. Statistical methods include the correlation coefficient method, principal component analysis method, Newton's interpolation method, nonlinear regression model, and so on. The accuracy obtained in this method is limited due to the inability to model nonlinear and multivariate data. Previous research conducted by S.J Horng [6] proposed predicting air quality with PM2.5 parameters using a deep learning model. Extensive experimental evaluation using two datasets, and the results show that the proposed model can handle PM2.5 air pollution forecasts with fairly good accuracy. Previous research was conducted by Khumaidi [7] to predict time series data based on air quality in the city of Bandung. The parameters used are PM10, ISPU, temperature, and humidity. Khumaidi uses LSTM modeling with 4 hidden layers; batch sizes are 32, using the adam optimizer, epoch worth 1000, and the mean squared error as a loss function. This shows that the model produces a fairly good predictive accuracy for three parameters (temperature, humidity, ISPU). This is indicated by the predicted RMSE value, which is smaller than the standard deviation value of the test dataset. However, the best prediction results from the four test parameters are humidity predictions followed by temperature, ISPU, and PM10.

Based on several previous studies, this study proposes a classification model based on pollution parameters that affect air quality based on ISPU data in DKI Jakarta using the deep learning method. ISPU is defined as a number that does not have a unit that describes the condition of ambient air quality in a particular location, which is based on the impact on human health, aesthetic value, and other living things [8]. This study focused on comparing the performance of the Long Short Term Memory (LSTM) and One Dimensional Convolutional Neural Network (1D CNN) algorithms to determine the best algorithm for predicting air quality. This research aims to provide information on suitable methods for classifying future air quality situations so that the government will consider the decision-making process to control air pollution and improve air quality in the future.

2. RESEARCH METHOD

In general, the method in this research will go through several stages, as shown in Figure 1. The stages in this research method are the framework used to facilitate the research process.

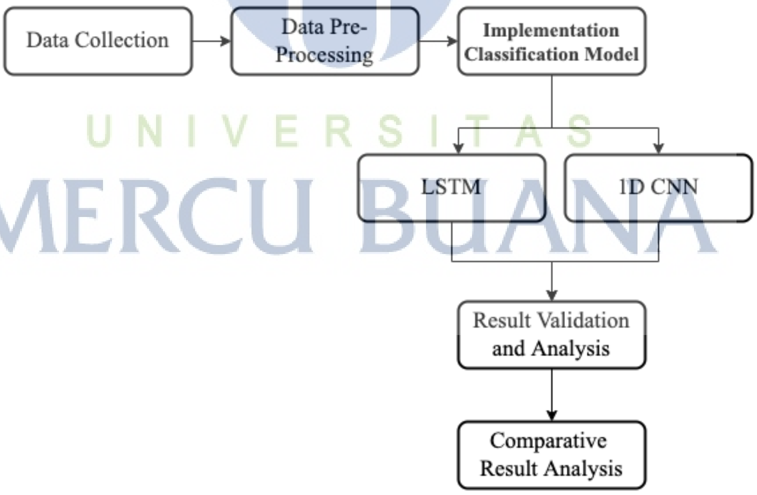

Fig. 1. The Proposed Research Methodology

2.1. Data Collection

The dataset used in this study is Air Pollution Standard Index (ISPU) data for five years from January 2016 to December 2020 obtained from Open Data Jakarta (site: https://data.jakarta.go.id). This dataset has 8376 rows of data with 10 columns obtained from the Automatic Ambient Air Quality Monitoring Station (SPKU) spread over five areas of DKI Jakarta. Parameters and data descriptions are shown in Table 1.

Table 1. Description of Dataset Parameters

Data	Description
Tanggal	Air quality measurement time

Stasiun	Measurement station location
PM10	Particulates measured for 24 hours
SO2	Sulfide (SO2 form) measured for 24 hours
CO	Carbon Monoksida yang diukur selama 8 jam
O3	Carbon Monoksida yang diukur selama 8 jam
NO2	Nitrogen Dioxide measured for 1 hour
Max	The highest measurement value of all parameters at the same time
Critical	Parameters with the highest measurement results
Categori	ISPU calculation result category

2.2. Data Preprocessing

After the data is collected, then the preprocessing stage is carried out. Preprocessing is concerned with preparing several data processed in a neural network. The dataset collected contains some missing values in some attributes [9]. Then delete some unneeded columns and transform the label data into numeric. In this study, we preprocessed the data as follows:

a. Data cleaning, cleaning data for noise found in the form of missing values [10], invalid data, or the same data in one column [11], [12]. The technique used to overcome the existing noise is by deleting the record.
b. Data selection, this process is the selection of attributes that will be used as a new dataset for processing [11]. This study's selection of attributes is based on the measured pollutant parameters, namely the air pollution standard index.
c. The data transformation changes the data type that was previously converted into an object type into a float type. Another transformation performs label encoding by changing class labels, which were previously of string type to a numeric type.
d. Data Normalization, To overcome inconsistent data, normalization is carried out so that the data distribution has an average of 0 and a standard deviation of 1 [13]. The method used for normalization is the sklearn StandardScaler library.

2.3. Implementation Classification Model

Data that has been preprocessed is divided into training data and test data. Training data is used to train the model, and test data is used to test the model. This study uses five parameters, including PM10, SO2, CO, O3, NO2, as input to be processed in LSTM and 1D CNN cells.

- **Long Short Term Memory (LSTM)**

  LSTM is an algorithm developed from the Recurrent Neural Network (RNN) unit to handle classification cases for long sequential data so that it can reduce vanishing gradients[9] and be able to learn on long-term and short-term dependencies [14]. LSTM generally has a memory cell, input gate, output gate, and forget gate. The input layer will receive feature data used as input, and then the feature enters the hidden layer. There will be three gates in the hidden layer, each connected through four interconnected layers. LSTM cell works by receiving input and storing it temporarily [15]. Then the cell status matrix and the current LSTM cell output will be obtained. With this method, the memory information has been stored and passed to the LSTM cell at the last point in time [16].

- **One Dimensional Convolutional Neural Network (1D CNN)**

  Convolutional Neural Network (CNN) is one of the most popular deep learning methods. CNN structure includes 1D CNN, 2D CNN, 3D CNN. CNN consists of a convolutional layer, a pooling layer, and a fully-connected layer [17]. The way 1D CNN works is that the convolutional layer is given data used as input. Then the convolutional and pooling layers extract features from the input. Each convolutional layer has a kernel of the same size, followed by the pooling layer doing the pooling method (average or maximum) and then sending the output to the fully-connected layer. The single hidden unit of a 1D CNN is only connected to the three inputs of the input layer. This kind of connection greatly reduces the number of parameters and speeds up the training process of the neural network.[11]–[13].

2.4. Validation and Analysis Results

In testing air quality on actual and predicted data, we use K-Fold cross-validation for the validation iteration process by randomly dividing the data into k parts with the same data size

[20]. Furthermore, the analysis of the performance and quality of the model is carried out by calculating the confusion matrix to get the values of accuracy, precision, recall, f1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{Precision\ x\ Recall}{Precision + Recall} \tag{4}$$

Explanation:

- TP: True Positive is the result of a correct classification that is classified as true
- TN: True Negative is the result of an incorrect classification that is classified as true
- FP: False Positive is the result of a true classification which is classified as false
- FN: False Negative is the result of an incorrect classification which is classified incorrectly

2.5. Experiment Scenarios

The test scenario is divided into 3 stages. The first test scenario is to test the classification model based on training and test data distribution. The second test scenario sets the hyperparameter values in the classification model based on the proportion of the best data previously obtained. The third test scenario is to validate using K-Fold Cross Validation with a different number of k. The flow chart of the test scenario is shown in Figure 2.

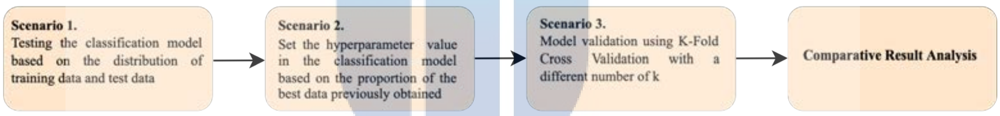

Fig. 2. The flow of experimental scenarios

3. RESULTS AND DISCUSSION

3.1. Data Preprocessing

After going through the stages of data collection through Open Data Jakarta, the pre-processing stage was carried out. The pre-processing stage begins with cleaning the data from empty values, data noise, and data duplication. After the data is clean, proceed with selecting attributes or parameters. The selection of attributes is based on the air pollution parameters in the Air Pollution Standards Index.

The data cleaning process is intended to clean unnecessary data or data that hinders the data mining process [1]. After data cleaning, the number of data was reduced to 6772 data. The results of the cleaning data can be seen in Table 2.

Table 2. Data After Cleaning

Tanggal	Stasiun	PM10	SO2	CO	O3	NO2	Max	Critical	Categori
2016-01-01	DKI1 (Bundaran HI)	59.0	19.0	27.0	31.0	1.0	59.0	PM10	SEDANG
2016-01-16	DKI1 (Bundaran HI)	59.0	42.0	26.0	95.0	4.0	95.0	O3	SEDANG
2016-01-17	DKI1 (Bundaran HI)	58.0	42.0	19.0	149.0	4.0	149.0	O3	TIDAK SEHAT
2016-01-18	DKI1 (Bundaran HI)	73.0	44.0	32.0	87.0	4.0	87.0	O3	SEDANG
2016-01-19	DKI1 (Bundaran HI)	60.0	45.0	24.0	78.0	4.0	78.0	O3	SEDANG

The next step is to select the data. The data selection aims to determine the attributes used as a new dataset for later data mining processes. The results of the data selection can be seen in Table 3.

Table 3. Data Selection Results

PM10	SO2	CO	O3	NO2
59.0	19.0	27.0	31.0	1.0
59.0	42.0	26.0	95.0	4.0
58.0	42.0	19.0	149.0	4.0
73.0	44.0	32.0	87.0	4.0
60.0	45.0	24.0	78.0	4.0

The next step is to normalize the data by setting the average data distribution to 0 and the standard deviation equal to 1 using the Sklearn Standardscaler library method. The results of normalization can be seen in Figure 3.

```
array([[ 0.39955512, -0.21927662,  0.52478944, -1.01104572, -0.86960769],
       [ 0.39955512,  1.68089303,  0.44705646,  0.81246586, -0.5906764 ],
       [ 0.33823417,  1.68089303, -0.0970744 ,  2.35105375, -0.5906764 ],
       ...,
       [-1.86932035,  0.93734839, -1.18533611, -0.89707624, -0.6836535 ],
       [-2.2372461 , -1.21066948, -1.34080207, -1.2959694 , -0.77663059],
       [-2.11460418, -0.71497305, -1.10760313, -1.2104923 , -0.6836535 ]])
```

Fig. 3. Hasil Normalisasi *StandardScaler*

3.2. Distribution of training data and testing data

This section presents the results of the accuracy of the experiments that have been carried out by applying the classification method using the Long Short Term Memory (LSTM) algorithm and the One Dimensional Convolutional Neural Network (1D CNN). Experimental scenario based on the distribution of training data and test data. The first experiment used 60% training data and 40% test data, and the second experiment used 70% training data and 30% test data; the third experiment used 80% training data and 20% test data. The values obtained using standard hyperparameters of each algorithm can be seen in Table 4 for LSTM and Table 5 for 1D CNN.

Table 4. LSTM Default Hyperparameter

Parameter	Nilai
Hidden unit LSTM	64, 128
Dense units	4
Optimizer	Adam
Learning rate	0.01
Loss function	Categorical Crossentropy
Epochs	100

Table 5. 1D CNN Default Hyperparameter

Parameter	Nilai
Filter convolutional layer	100, 50
Kernel filter	1
Max Pooling layer size	1
Dense units	4
Learning rate	0.01
Loss function	Categorical Crossentropy
Optimizer	Adam
Epochs	100

Table 6 presents the results of the performance of the LSTM model using the train test split, while Table 7 presents the results of the CNN 1D performance using the train test split. Based on the best level of accuracy for the two algorithms in this scenario, the accuracy of the LSTM is

97.93% at the proportion of 80% train set and 20% test set (see Table 6), while the 1D CNN accuracy is 90.23% with 80% train set and 20% test set (see Table 7). This shows that the LSTM and 1D CNN models require optimal training data for better classification accuracy and minimizing overfitting and underfitting problems.

Table 6. LSTM Performance is Based on Train Test Split

Experiment	LSTM			
	Accuracy	**Precision**	**Recall**	**F1-Score**
60/40	97.67%	97.24%	83.38%	87.07%
70/30	97.69%	97.23%	80.49%	83.88%
80/20	97.93%	97.52%	93.81%	95.43%

Table 7. 1D CNN Performance is Based on Train Test Split

Experiment	1D CNN			
	Accuracy	**Precision**	**Recall**	**F1-Score**
60/40	90.03%	90.80%	81.15%	84.83%
70/30	88.39%	91.86%	66.19%	71.88%
80/20	90.23%	82.10%	91.30%	85.70%

3.3. Hyperparameter tuning

The second experiment aims to perform several hyperparameter settings to optimize and improve algorithm performance. The hyperparameters performed by the tuning process include the activation function, learning rate, batch size, and epoch. Obtained from a sequential comparison process by trying all combinations of hyperparameter values, from the experimental results obtained the optimal hyperparameter for the LSTM with the activation function = tanh, learning_rate = 0.01, batch_size = 64, and epochs = 100. As for the CNN 1D hyperparameter, the optimal functions are activation=relu, learning rate=0.001, batch_size=64, and epochs=100.

Furthermore, the test results before and after hyperparameter tuning are presented in Tables 8 and 9. Table 8 shows the LSTM evaluation values before and after parameter setting; the accuracy score increased by 0.31% from 98.73% to 99.04%. Table 9 shows the evaluation value of 1D CNN before and after parameter setting, with the accuracy score increasing from 89.53% to 98.32% or 8.79%. These results indicate that setting hyperparameters can affect the performance of both algorithms.

Table 8. LSTM Results Before and After Parameter Tuning

Evaluation Score	LSTM	
	Non Hyperparameter Tuning	**Hyperparameter Tuning**
Accuracy	98.73%	99.04%
Precision	98.92%	99.38%
Recall	96.98%	98.66%
F1-Score	98.73%	99.01%

Table 9. 1D CNN Results Before and After Parameter Tuning

Evaluation Score	1D CNN	
	Non Hyperparameter Tuning	**Hyperparameter Tuning**
Accuracy	89.53%	98.32%
Precision	81.01%	97.20%
Recall	89.51%	90.24%
F1-Score	83.92%	93.09%

3.4. Result Validation and Analysis

We carried out further experiments with the cross-validation method from the previous model scenario to validate the above results. We performed three k-fold different cross-validation scenarios, including 5-fold, 10-fold, 20-fold [20] with the proportion of data in the previous scenario applied to LSTM and 1D CNN. The results can be seen in Tables 10 and 11.

Table 10. LSTM Cross-Validation Results Performance Testing

Cross Validation	LSTM				
	Accuracy	Precision	Recall	F1-Score	Time(s)
5-fold	98.67%	98.64%	98.85%	98.74%	315.049s
10-fold	98.45%	88.06%	97.96%	91.41%	681.038s
20-fold	98.30%	98.16%	88.65%	92.16%	1363.489s

Table 11. 1D CNN Cross-Validation Results Performance Testing

Cross Validation	1D CNN				
	Accuracy	Precision	Recall	F1-Score	Time(s)
5-fold	97.71%	97.96%	82.85%	87.19%	104.531s
10-fold	97.79%	98.19%	87.83%	91.76%	204.462s
20-fold	98.08%	98.29%	93.30%	95.51%	416.74s

The cross-validation results above show that the LSTM algorithm obtains the best accuracy of 98.67% on a 5-fold cross-validation with an execution time of 315,049s. For comparison, the CNN 1D algorithm obtained the best accuracy of 98.08% with a time of 416.74s. This shows that LSTM provides better accuracy values for fewer k-folds. In comparison, CNN 1D obtained a better accuracy value at a larger k-fold [20].

3.5. Comparative Result Analysis

Of all the test scenarios that have been carried out with the results of the classification of each model based on experiments that have been carried out using air pollution parameters including PM10, SO2, CO, O3, NO2, providing several comparisons results. In the first scenario comparing the distribution of training data and test data, the best data distribution is obtained at the proportion of 80% training data and 20% test data, wherein this test LSTM obtains an accuracy of 97.93% and 1D CNN obtains an accuracy of 90.23% (see Tables 2 and 3). The second test scenario is to do some hyperparameter settings to find the best accuracy value. The results of this scenario show that changing some hyperparameters can improve the performance of both algorithms, where the LSTM accuracy increases by 0.31% (see Table 5) and 1D CNN accuracy increases by 8.79% (see Table 6). Furthermore, to ensure the results of the two algorithms, validation is carried out, which is presented in Tables 7 and 8. There is a clear difference between the classification results using the LSTM and CNN 1D algorithms. Judging from all the tests, the LSTM model shows better results than the CNN 1D model because the LSTM has a more complex gate to learn short-term and long-term information on the data. If we look at the execution time, 1D CNN is faster than the LSTM model execution time, and this is because 1D CNN has fewer hidden unit connections in the input layer, this can speed up the neural network training process [17], [21], [22].

4. CONCLUSION

Several comparisons have been made from all experimental scenarios that have been carried out by implementing, analyzing, and validating the results. To get the best classification model with pollution parameters that affect air quality, the comparison of LSTM and 1D CNN provides a significant performance. The data used in this study is ISPU data in DKI Jakarta. The data used goes through the pre-processing process first to clean the data and select the attributes used in the data mining process. The results showed that the LSTM algorithm obtained an accuracy of 98.67% on a 5-fold cross-validation with an execution time of 315,049s. For comparison, the CNN 1D algorithm obtained an accuracy of 98.08% on 20-fold cross-validation with a time of 416.74s. LSTM provides better accuracy values for fewer k-folds.

In comparison, CNN 1D obtains a better accuracy value at a larger k-fold. Judging from all the tests, the LSTM model shows better results than the CNN 1D model because the LSTM has a more complex gate to learn short-term and long-term information on the data. When viewed from the execution time, CNN 1D is faster than the LSTM model execution time because 1D CNN has fewer hidden unit connections in the input layer, which can speed up the neural network training process. It was concluded, with the type of quantitative data and the characteristics of a low level of data variation, LSTM could be the proper method in classifying air quality. It could be an alternative to assist the government in handling air pollution so that people can enjoy healthy air.

Acknowledgments

We would like to thank colleagues at the Faculty of Computer Science, Universitas Mercu Buana for their support.

5. REFERENCES

[1] L. D. Wijaya, "Kualitas Udara Dunia 2020: Polusi di Jakarta Tetap Tinggi Selama PSBB," *https://metro.tempo.co/read/1443299/kualitas-udara-dunia-2020-polusi-di-jakarta-tetap-tinggi-selama-psbb/full&view=ok*, Mar. 18, 2021.

[2] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, pp. 1–44, 2018, doi: 10.3390/ijerph15040780.

[3] Q. Tao, F. Liu, Y. Li, and D. Sidorov, "Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU," *IEEE Access*, vol. 7, pp. 76690–76698, 2019, doi: 10.1109/ACCESS.2019.2921578.

[4] M. G. Ragab *et al.*, "A novel one-dimensional cnn with exponential adaptive gradients for air pollution index prediction," *Sustainability (Switzerland)*, vol. 12, no. 23, pp. 1–22, 2020, doi: 10.3390/su122310090.

[5] T. C. Bui, V. D. Le, and S. K. Cha, "A deep learning approach for forecasting air pollution in south korea using LSTM," *arXiv*, 2018.

[6] S. Du, T. Li, Y. Yang, and S. J. Horng, "Deep Air Quality Forecasting Using Hybrid Deep Learning Framework," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2412–2424, Jun. 2021, doi: 10.1109/TKDE.2019.2954510.

[7] A. Khumaidi, R. Raafi, I. Permana Solihin, and J. Rs Fatmawati, "Pengujian Algoritma Long Short Term Memory untuk Prediksi Kualitas Udara dan Suhu Kota Bandung," *Jurnal Telematika*, vol. 15, no. 1.

[8] A. Kurniawan, "PENGUKURAN PARAMETER KUALITAS UDARA (CO, NO2, SO2, O3 DAN PM10) DI BUKIT KOTOTABANG BERBASIS ISPU," *Jurnal Teknosains*, vol. 7, no. 1, p. 1, Jul. 2018, doi: 10.22146/teknosains.34658.

[9] Y. Chang, H. Chiao, S. Abimannan, Y. Huang, Y. Tsai, and K. Lin, "An LSTM-based aggregated model for air pollution forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451–1463, 2020, doi: 10.1016/j.apr.2020.05.015.

[10] Y. W. Romadon and D. Fitrianah, "The comparative study on clustering method using hospital facility data in Jakarta District and surrounding areas," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, pp. 749–755, 2019, doi: 10.18178/ijmlc.2019.9.6.868.

[11] D. Lazuardi Aditya and D. Fitrianah, "COMPARATIVE STUDY OF FUZZY C-MEANS AND K-MEANS ALGORITHM FOR GROUPING CUSTOMER POTENTIAL IN BRAND LIMBACKid 2 (*) Corresponding Author." [Online]. Available: https://www.mercubuana.ac.id

[12] L. Nurellisa and D. Fitrianah, "Analisis Rekomendasi Calon Debitur Motor Pada Pt . Xyz Analysis of Motorcycle Debitor Recommendations in Pt . Xyz Using," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 4, pp. 673–682, 2020, doi: 10.25126/jtiik.202072080.

[13] A. Oktaviani and Hustinawati, "PREDIKSI RATA-RATA ZAT BERBAHAYA DI DKI JAKARTA BERDASARKAN INDEKS STANDAR PENCEMAR UDARA MENGGUNAKAN METODE LONG SHORT-TERM MEMORY," *Jurnal Ilmiah Informatika Komputer*, vol. 26, no. 1, pp. 41–55, 2021, doi: 10.35760/ik.2021.v26i1.3702.

[14] M. F. Maulana, S. Sa'adah, and P. Eko Yunanto, "Crude Oil Price Forecasting Using Long Short-Term Memory," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 7, no. 2, p. 286, Sep. 2021, doi: 10.26555/jiteki.v7i2.21086.

[15] A. Khumaidi, R. Raafi'udin, and I. P. Solihin, "Pengujian Algoritma Long Short Term Memory untuk Prediksi Kualitas Udara dan Suhu Kota Bandung," *Jurnal Telematika*, vol. 15, no. 1, pp. 13–18, 2020, Accessed: Nov. 01, 2021. [Online]. Available: https://journal.ithb.ac.id/telematika/article/view/340

[16] Y. S. Chang, H. T. Chiao, S. Abimannan, Y. P. Huang, Y. T. Tsai, and K. M. Lin, "An LSTM-based aggregated model for air pollution forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451–1463, Aug. 2020, doi: 10.1016/j.apr.2020.05.015.

[17] C. Lang, F. Steinborn, O. Steffens, and E. W. Lang, "Electricity Load Forecasting -- An Evaluation of Simple 1D-CNN Network Structures," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.11536

[18] M. G. Ragab *et al.*, "A novel one-dimensional cnn with exponential adaptive gradients for air pollution index prediction," *Sustainability (Switzerland)*, vol. 12, no. 23, pp. 1–22, Dec. 2020, doi: 10.3390/su122310090.

[19] H. Cheng, Z. Xie, Y. Shi, and N. Xiong, "Multi-Step Data Prediction in Wireless Sensor Networks Based on One-Dimensional CNN and Bidirectional LSTM," *IEEE Access*, vol. 7, pp. 117883–117896, 2019, doi: 10.1109/ACCESS.2019.2937098.
[20] M. Sadikin and F. Alfiandi, "Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, p. 4763, Dec. 2018, doi: 10.11591/ijece.v8i6.pp4763-4771.
[21] W. Tang, G. Long, L. Liu, T. Zhou, J. Jiang, and M. Blumenstein, "Rethinking 1D-CNN for time series classification : A stronger baseline".
[22] M. Abo-Tabik, N. Costen, J. Darby, and Y. Benn, "Towards a smart smoking cessation app: A 1D-CNN model predicting smoking events," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–18, 2020, doi: 10.3390/s20041099.

BIOGRAPHY OF AUTHORS

	**Jodikal Pomalingo** is a undergraduate student in Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia. He was born in Gorontalo on February 19th, 2000. He is interested in data mining, mobile programming and web programming. Her current research scope is data mining, machine learning. Email: jodikalpomalingo@gmail.com
	**Devi Fitrianah** She received the Bachelor's degree in Computer Science from Bina Nusantara University, Jakarta, Indonesia, in 2000, and the Master's degree in Information Technology and Ph.D. degree in Computer Science from the Universitas Indonesia, Depok, Indonesia, in 2008 and 2015, respectively. In 2014, she had a sandwich program at the Laboratory for Pattern Recognition and Image Processing and GIS (PRIPGIS Lab) Department of Computer Science, Michigan State University, East Lansing, Michigan, USA. She is currently a Faculty Member with the Department of Computer Science, Universitas Mercu Buana. Her research interests include machine learning, data mining, applied remote sensing, and geographic information system. Email: devi.fitrianah@mercubuana.ac.id

KERTAS KERJA

Ringkasan

Kertas kerja ini berisi penjelasan mengenai kelengkapan material yang digunakan dan semua hasil penelitian Tugas Akhir yang tidak dimuat dalam artikel jurnal yang berjudul "*Comparative Study of Deep Learning Methods LSTM and 1D CNN Algorithm: Case Study of Air Pollution Standard Index Data in DKI Jakarta*". Dalam kertas kerja ini menyajikan beberapa bagian yang dimulai dengan literatur *review*, analisis perancangan, *source code*, informasi himpunan data yang digunakan, tahapan eksperimen dan seluruh hasil eksperimen.

- Bagian 1: Literatur *review* menjabarkan mengenai beberapa referensi artikel jurnal yang terkait dengan penelitian.
- Bagian 2: Analisis dan Perancangan membahas analisis masalah dan analisis model pada penelitian ini. Pada analisis masalah berisi mengenai latar belakang penelitian, sedangkan analisis model berisi rancangan diagram yang digunakan.
- Bagian 3: *Source code* berisi kumpulan kode pada setiap proses mulai dari membaca data, data *Preprocessing*, implementasi model, pengujian dan validasi.
- Bagian 4: *Dataset* melampirkan informasi tentang *dataset*
- Bagian 5: Tahap eksperimen menjelaskan tahapan dalam eksperimen yang dimulai dari pengumpulan data, data *preprocessing*, implementasi model, pengujian dan validasi model.
- Bagian 6: Hasil implementasi eksperimen secara keseluruhan yang telah dilakukan seperti yang disebutkan pada Bagian 5. Bagian ini berisi validasi dan analisis dari setiap model yang kemudian dilakukan analisis hasil perbandingan untuk menemukan model klasifikasi berdasarkan parameter polusi udara.