JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 2, Juni 2024, Pp. 476-486

COMPARISON OF KNN AND LSTM ON THE PREDICTION OF THE OPERATIONAL CONDITIONS OF NATURAL GAS PIPELINE TRANSMISSION NETWORKS

Afrizal Syahruluddin Yusuf[1], Hasmawati[2], Aditya Firman Ihsan[3]

1. Telkom University, Indonesia
2. Telkom University, Indonesia
3. Telkom University, Indonesia

Article Info

Keywords: forecasting; KNN; LSTM; pressure; pipeline transmission; natural gas

Article history:
Received 4 February 2024
Revised 18 February 2024
Accepted 3 March 2024
Available online 1 June 2024

DOI :
https://doi.org/10.29100/jipi.v9i2.4528

Corresponding Author
Afrizal Syahruluddin Yusuf
E-mail address:
afrizalsyahruluddinyusuf@gmail.com

ABSTRACT

During the gas distribution process, a sequence of compressors creates a pressure difference, causing gas to move from regions of high pressure to areas with comparatively lower pressure. The Natural Gas transmission process experiences variations in pressure and temperature, primarily caused by frictional losses, differences in altitude, gas velocity, and the Joule-Thompson effect. Additionally, effective heat transfer to or from the environment contributes to temperature changes throughout the pipeline. The presence of liquid and density changes (hydrate) within the channel also has an impact on the pressure, influencing both pressure and temperature conditions. This study implements the KNN and LSTM models to predict pressure conditions in natural gas transmission pipelines to analyze the performance comparison of the best model performance using several appropriate parameters to support maximum method performance results. The results show that the LSTM model is better at predicting pressure conditions in natural gas pipeline transmission networks, with an R^2 score of 99.45, compared to the KNN model, with an R^2 score of 92.82. This study also obtained prediction results from the KNN and LSTM models; the KNN model tends to produce the same pressure value for eight months, while the LSTM model produces pressure values that tend to vary.

I. INTRODUCTION

THE oil and natural gas industry is a potential industry indicator of a country's economic growth. In the gas distribution process involving a sequence of compressors generating pressure differences, gas moves from a region of higher pressure to one with comparatively lower pressure. These compressors, driven by electric or natural gas engines, compress or squeeze the incoming gas, propelling it at higher pressure. Typically, compressors used for larger transmission lines are significantly larger than those employed for transporting gas through smaller distribution lines to homes. However, certain gas collection systems do not need a compressor, as the natural pressure from the well is enough to move the gas through the collection ducts. [1].

In the Natural Gas transmission process, variations in pressure and temperature occur due to factors such as frictional losses, differences in altitude, gas velocity, and the Joule-Thompson effect. Additionally, efficient heat transfer with the surroundings causes temperature fluctuations along the pipeline. Alterations in pressure and temperature, along with the formation of liquid and density changes (hydrate), also impact the overall pressure in the channel [1]. With every change that occurs, the composition of natural gas can change, so changes in pressure at each distribution point must be controlled continuously so that this does not happen [2].

In the context of modelling and prediction in this industry, several studies have been conducted, such as in a journal entitled " Implementasi Weighted K-Nearest Neighbor Untuk Peramalan Data Deret Waktu", stating that the best model for predicting the Gross Domestic Product (GDP) of forestry and logging. The model has undergone several validation processes to ensure reliable and accurate predictions, with k = 13 and a MAPE value of 0.0038% [3]. In addition, research entitled" Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk PT. Metiska Farma" explained that the LSTM model produces an average percentage of model error between the predicted value and the smallest actual value per day using MAPE is 12% [4].

Kurniawan C.'s research entitled "Penerapan Metode KNN-Regresi dan Multiplicative Decomposition untuk Prediksi Data Penjualan pada Supermarket X " evaluates the accuracy of two KNN methods, namely Regression KNN and Multiplicative Decomposition KNN. The results showed that the Regression KNN has an average Root Mean Squared Error (RMSE) of 757.77 and a MAPE of 0.36, while the Multiplicative Decomposition KNN has an average RMSE of 492.89 and a MAPE of 0.29. Both methods provide fairly accurate prediction results for sales data at supermarket X [5]. Another study by Martinez F. in the journal "Time Series Forecasting with KNN in R: the tsfknn Package" resulted in an evaluation of the KNN model for industrial data. The results showed that this KNN model resulted in an RMSE of 274.19569, a Mean Absolute Error (MAE) of 202.69048, and a MAPE of 11.09727. This evaluation indicates that the KNN model is quite effective in predicting time series data for the industrial data tested [6].

The study, " Metode K-Nearest Neighbor Untuk Memprediksi Penjualan Produk Pada UMKM Pengolahan Ikan Maju Jaya," by Izzan A. compared 13 KNN methods to predict sales of processed tekwan products. The results show that the selected KNN method produces an RMSE of 3,234. This study shows the potential of the KNN method to predict product sales in micro, small and medium enterprises (MSMEs) with satisfactory results [7]. Then there is also research by Wang Z. entitled "Short-Term Traffic Volume Forecasting with Asymmetric Loss Based on Enhanced KNN Method" focuses on predicting short-term traffic volume using Enhanced KNN. The evaluation results show that using the normal model with smoothed data produces a Mean Squared Error (MSE) of 124.30, a Mean Absolute Error (MAE) of 7.74, and an Integrated Mean Squared Error (IMSE) of 138.77. This study shows the potential of the KNN model in overcoming traffic volume predictions with an emphasis on asymmetric errors [8].

Studies related to this research were previously conducted in a journal entitled "Multi-Layer LSTM Implementation in Operational Condition Forecasting of a Natural Gas Transmission Pipeline Network" by Firman A. This study solves the problem of accurate transmission monitoring on gas pipelines, intending to predict pressure and flow in gas pipelines by the two combined methods. This research combines Recurrent Neural Network (RNN) and Long Term Memory Cell (LSTM) methods. In this study, the method produced good pressure predictions with a performance error of 2%. Meanwhile, the estimated flow in the pipeline fluctuates very high. Therefore, the pressure prediction obtained from this method is more accurate than the flow prediction obtained [9]. In the following year's study by Patra S. entitled "Regional Groundwater Sequential Forecasting Using Global and Local LSTM Models" shows the best results on the t+1 prediction with an RMSE value of 0.557, MAE of 0.458, and R Squared of 0.938. These results indicate that the LSTM method effectively predicts time series data for increasing groundwater levels in certain areas by predicting one-time steps [10].

The latest study by Bayram F. entitled "DA-LSTM: A Dynamic Drift-Adaptive Learning Framework For Interval Load Forecasting With LSTM Networks" discusses the prediction of electrical loads using the Dynamic Drift-Adaptive Learning Framework with the LSTM model. The results show that this model has the best performance with a MAPE of 8.26% and an RMSE of 21.41% when tested using the local CPU and GPU on the device and the CPU and GPU on the cloud-web service (AWS). This study demonstrates the potential and reliability of the LSTM model for use in predicting electrical loads [11].

Based on this background, a model was built to predict the pressure condition of the natural gas pipeline transmission network using the KNN and LSTM models. Research that specifically examines comparing the KNN and LSTM models to predict pressure conditions in natural gas pipeline transmission is not yet available, so it is still interesting to do further research. The model is used to predict pressure conditions in natural gas transmission pipelines. By using the KNN and LSTM models, it is possible to compare the performance of the two models in predicting pressure conditions in natural gas pipeline transmission. Then, with the modeling built, it can be seen how the performance of the best model is compared by using the right parameters to support the maximum method performance results.

The limitation of this study is that the authors only use records of pressure data from oil and gas company gas pipeline transmission operations located in the Natuna Sea with a total of not more than 9000 data per hour in the time range from August 2020 to July 2021. This final project aims to find out the results of implementing the KNN and LSTM models to predict pressure conditions in natural gas pipeline transmission. In addition, it also compares the performance of the KNN and LSTM models to predict pressure conditions in natural gas pipeline transmission with R Squared, RMSE, MAPE, and MAE metrics.

II. Research Methodology

The dataset used in this final project results from data records from operational pressure data for oil and gas pipeline transmissions in the Natuna Sea. After the dataset has been successfully collected, it enters the data pre-processing stage, which is divided into train, test, and predictive data. Train data is used to train the model, test data is used to test the model for tuning or optimization of the model, and predictive data is predictive data from models that have been previously trained and tested. In this final project using the KNN and LSTM models, the final step is to evaluate the results obtained after the model has been successfully built.

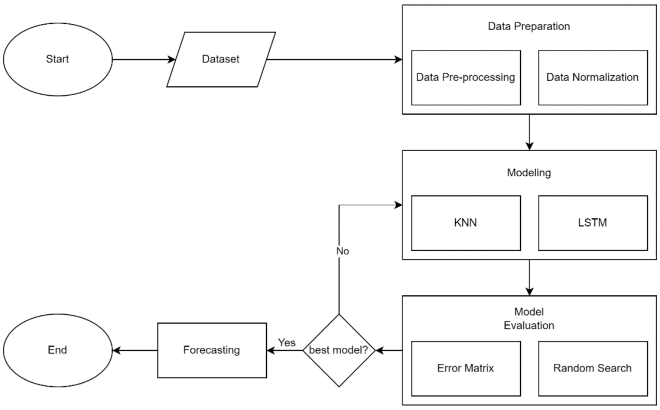

Fig. 1. System Design Flowchart

A. Dataset

The dataset used in this study as input for the KNN and LSTM models applied results from data records from operational data transmission of oil and gas companies' gas pipelines located in the Natuna Sea [12], [13]. Equipment at the natural gas distribution facility sends many data on an ongoing basis on the physical state of the gas itself, such as pressure and temperature. In addition to the physical state of the gas, facilities on the gas distribution equipment also send data in the form of the composition of the gas, such as example ethane, nitrogen, carbon dioxide and water. So overall, the dataset has 17 features, with 4 being the physical state of the gas being distributed and 13 features of the gas composition. The gas is a hydrate, which is also a chemical compound containing water molecules that can normally be removed by heating to a certain temperature. The dataset that will be used is a dataset obtained from August 2020 to July 2021. Data collection took place between August 2020 and July 2021, with a total of 21 attributes consisting of distribution time, gas sending source, pressure, temperature, and condition and composition of the gas then what is used is the pressure attribute data per hour which then enters the data preparation stage.

B. Preprocessing Data

This study used data for one year with a total of 61313 data lines. In the early stages of data pre-processing, data was first prepared and analyzed before being implemented into the model and in the dataset, "ASSET_ID" was selected with the code number "133060", for example, as the source gas, which will later be distributed or can be referred to as an inlet point. Natural gas pressure data goes through a pre-processing stage before being used in training and testing the KNN and LSTM models. After selecting the gas delivery source or inlet point, the data pre-processing stage includes data normalization by scaling the pressure features using the min-max scaler method so that all features have the same scale with the average shifting to zero and the standard deviation becoming one and missing handling. Values so that the total data processed for model design is 8759 lines of data. Then the data is divided into two parts, namely train data and test data, with a proportion of 70:30.

C. K-Nearest Neighbor (KNN)

The KNN model is a method that implements a supervised algorithm. The primary objective of supervised learning is to detect novel patterns, whereas unsupervised learning is focused on identifying patterns within the given data. One advantage of KNN is that it can be used for any data. Numerical or non-numerical. Discrete or continuous. KNN Regression is used to predict the output value of data. The working principle of the K-Nearest Neighbor (KNN) model includes determining the shortest distance between the data that was carried out before the evaluation stage with the K closest neighbours from the training data. Train data is projected onto a multidimensional space, where each dimension describes a property of the data. The space is subdivided based on classification in the form of data sections. The KNN training process produces a value of k which is expected to provide the highest performance to generalize future data. However, until now, k cannot be determined mathematically. No formula can be used. The only way is trial and error, so the training process observes several k until the optimum k is produced.

KNN also has several areas that need attention, namely being sensitive to less relevant features. To overcome the weaknesses of KNN, experts have proposed various improvements which are generally proven to increase the accuracy of KNN significantly. Until now, dozens of KNN variants have various improvements based on different concepts and schemes. One of these improvements is improving the distance function to reduce the sensitivity of KNN to less relevant features. However, a KNN with high accuracy but impractical in terms of time and memory would be well-spent. Therefore, it is necessary to improve data structures to reduce time and memory complexity.

Many data structures have advantages and disadvantages, which can be selected according to the data set and model being built, for example, Ball-Tree, kd-Tree, and B-Tree. These data structures are designed to reduce the time complexity of finding the k-nearest neighbours [14].

D. Long Short-Term Memory (LSTM)

According to Heizer and Render (2005), the prediction time range can be grouped into three categories, including the following [15]:

a. Short-Term prediction for less than three months.
b. Medium-term (Medium-Term) predictions for three months to three years.
c. Long-Term predictions for more than three years.

One type of Recurrent Neural Network (RNN), namely LSTM, was introduced by Hochreiter and Scmidhuber in 1997, aiming to eliminate the weaknesses of RNN. The Long Short-Term Memory (LSTM) model, belonging to the category of recurrent neural networks (RNNs), is extensively employed in a diverse range of data sequence prediction and modeling tasks. LSTMs are specifically designed to address a problem in traditional RNNs related to the vanishing gradient problem, where long-term information tends to be lost as the sequence depth of the processed data increases. This makes LSTMs very effective at processing sequential data with complex temporal dependencies.

The LSTM model has a special mechanism called the "cell state," which acts as the model's long-term memory. The cell state allows the model to remotely remember important information in data sequences and retain that information through multiple timestamps. Furthermore, LSTM incorporates three gates—the forget gate, input gate, and output gate—that play a pivotal role in controlling the influx and efflux of information within the cell state. These gates effectively manage the flow of data, contributing to LSTM's ability to process and manage information in a sophisticated manner. These gates allow the model to control how much information to remember or forget and how much information to use to generate predictions.

LSTM's ability to handle long-term temporal dependencies and remember complex patterns in data sequences makes them useful in many prediction applications, such as stock price prediction, continuous text sentiment analysis, and natural language modelling. Using LSTM, models can capture important information in data history and provide more accurate predictions for future data. Although LSTM has higher complexity than other models, such as feedforward models, its unique ability to deal with sequential data makes it a strong choice in various prediction tasks. LSTM possesses the capability to discern the data that should be retained and the data that should be disregarded due to its operational mechanism, which relies on multiple gates, encompassing input gates, output gates, and forget gates. Through these distinct gates, LSTM effectively learns to store pertinent information while discarding irrelevant data, enabling it to process and retain essential patterns effectively. [16].

E. Random Search

This study needs tuning on the model, which aims to determine which parameters influence the model being tested. One method to find the optimal parameter combination is the random search method. Random search hyperparameters represent a technique utilized for optimizing machine learning models, aimed at discovering the most effective combination of hyperparameters. These hyperparameters, distinct from those learned by the model during training, necessitate prior determination by the user or researcher. Instances of hyperparameters found within machine learning models comprising the learning rate, denoting the magnitude of the step taken while employing gradient descent for optimization, the quantity of neurons housed within the concealed layer, governing the model's capacity to grasp intricate patterns, and the number of epochs, signifying the total iterations made over the complete training dataset, ultimately impacting the model's convergence and generalization abilities.

In the hyperparameter random search method, each hyperparameter is taken from a predetermined random distribution, such as a uniform or normal distribution. For example, suppose the model has two hyperparameters, such as the learning rate and number of neurons. Each hyperparameter will be drawn from a random distribution that fits within a predetermined range of values. Then, the model will be trained with a combination of these hyperparameters, and the performance of the model will be measured using relevant evaluation metrics, such as accuracy or mean squared error.

The advantage of random hyperparameter search is its ability to explore the hyperparameter space randomly and efficiently, making it possible to find the optimal combination relatively quickly. This method is also easier to implement and more scalable than the grid search method, where all hyperparameter combinations must be tested individually. However, remember that the random search method does not guarantee to find the best combination of hyperparameters but can often produce good results with more efficient efforts.

F. Model Training

After the data pre-processing, the next step is model training. Model training is one of the main stages in machine learning which aims to teach an algorithm or computer model to recognize complex patterns in data and make accurate predictions. This process is carried out by providing examples of data with known results to the model, and then the model will learn to find relationships and patterns among the data. The ultimate goal of model training is for the model to generalize and provide the right predictions for data that has never been seen before. The model training begins by initializing the model with random parameters or certain initial values. Then, the prepared data will be divided into two main parts: training data and validation data. Training data is used to teach the model, while validation data measures how well the model performs on data it has never seen.

During training, the model will calculate predictions for each training data example and compare them with actual results. Then, the model will correct itself using an optimization algorithm to reduce prediction errors. This process occurs repeatedly in the hope that the model will get closer to the actual results and eventually achieve a good performance on validation data. After the model has been trained and has reached a satisfactory level of performance, the model can be used to make predictions on new data that it has never seen before. In this final project, two models are used, namely KNN and LSTM. The train data used is pressure data on natural gas pipeline transmission. The training model is used to obtain a model that proceeds to the model testing stage.

G. Model Testing

Model testing is an important stage in machine learning that is carried out after the model has been trained with training data and validated to ensure that the model performs well on data it has never seen before. At this stage, the model will be tested using test data separate from the training and validation data. This test data is a representation of the real-world situation that the model wants to predict. This test aims to measure the model's generalizability, namely the extent to which the model can provide accurate predictions on new data. During the model testing process, test data will be given to the model, and the model will generate predictions based on the information learned during the training process. The predicted results will then be compared with the labels or actual values in the test data. Model performance evaluation uses metrics such as R Squared, MAPE, MAE, and RMSE, or specific metrics depending on the type of problem encountered.

The main goal of model testing is to objectively measure model performance and get an idea of how well the model performs on real-world data. If the model has performed satisfactorily on test data it has never seen, it can be used for predictive tasks in production environments or implementation in real applications. Suppose the model's performance is still unsatisfactory. In that case, it may be necessary to fine-tune the model or return to the training stage to improve the quality of the model before using it in critical situations. After the testing is successfully carried out, then the output of the process is produced, and it is continued in the evaluation process to find out the results obtained from the testing process.

H. Test Scheme

The models used in this study are KNN and LSTM, whose performance is compared by measuring the metrics R Squared, MAPE, MAE, and RMSE. Alone produces an effect on the model tested [17]. One method to find the optimal parameter combination is the random search method. Random search searches for combinations of parameters in a predetermined range of values so that testing of combinations of parameters is easy to do [17]. The KNN model parameters are neighbors, weights, and algorithms, while the LSTM model parameters that can be tuned include dropout rates, number of units, dense units, and dense activation [17]. The KNN model is used to process pressure time series data on gas to obtain linear predictions and the best parameters [18]. Meanwhile, the LSTM model is used to analyze further and predict residual nonlinearity in the data that affects the pressure time series on the gas and get the best parameters [19], [20].

I. Model Analysis and Evaluation

After the dataset has been integrated into the model, the next step involves the model evaluation phase. During this stage, various metrics are used to assess the performance of the model results after applying the dataset. This metric includes the first value R Squared, the second MAE, the third MAPE, and finally, the RMSE, which allows a comprehensive evaluation of the effectiveness and accuracy of the model. The results of each model will be measured for its performance using the metrics mentioned above. Suppose the error value for the model is too high. In that case, it can be said that it is not optimal because this can be caused by underfitting or overfitting the model, so the model implementation process is repeated [3].

Evaluation metrics in this study include the first R Squared, the second MAE, the third MAPE, and the last RMSE. R Squared (R^2), ranging from 0 to 1, signifies the degree to which the collective impact of independent variables influences the dependent variable's value. It is employed to assess the influence of specific independent latent variables on the dependent latent variable. RMSE, on the other hand, is an alternative method to evaluate prediction techniques, measuring the accuracy of the model's prediction results that have been tested. [18]. RMSE represents the squared average of errors in the model and is a readily applicable metric in prediction-related studies [19]. MAPE calculates the average absolute discrepancy between the predicted and actual values, represented as a percentage of the actual value. MAE calculates the mean absolute discrepancy between the predicted and actual values without incorporating any percentage scaling. The R Squared, RMSE, MAPE, and MAE formulas are as follows [21], [22]:

1) R Squared Score (R^2 Score)

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1}$$

Where R^2 is the coefficient of determination, RSS is the sum of the squares of the remainder, and TSS is the total sum of the squares.

2) Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(\tilde{y} - y_i)^2} \tag{2}$$

The predicted value, denoted $\tilde{y}_i$ s compared with the actual value, y_i, in the dataset, with n representing the total data points available.

3) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\tilde{y} - y_i| \tag{3}$$

The predicted value, denoted $\tilde{y}_i$ s compared with the actual value, y_i, in the dataset, with n representing the total data points available.

4) Mean Absolute Percentage Error (MAPE)

$$MAPE = 100 \times \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\tilde{y} - y_i}{y_i}\right| \tag{4}$$

The predicted value, denoted $\tilde{y}_i$ s compared with the actual value, y_i, in the dataset, with n representing the total data points available.

III. Result and Discussion

This section will explain the implementation of KNN and LSTM in predicting pressure conditions in natural gas pipeline transmission networks. The explanation in this section will be divided into two: the test results and the analysis of the test results. The test results will discuss the data that entered the testing phase to the random search process for hyperparameter combinations. Meanwhile, in analyzing the test results, an analysis process is carried out to determine which model is better. It has the maximum performance in predicting the pressure conditions of the natural gas pipeline transmission network and the results of pressure prediction in the KNN and LSTM prediction models.

A. Dataset

The dataset is obtained from data records from operational pressure data for oil and gas pipeline transmissions located in the Natuna Sea. An example of the dataset obtained can be seen in Table I. The total dataset obtained is 61313 data from company data records. After going through the pre-processing stages, the total dataset becomes 8759 because it uses "ASSET_ID" with the code "133060," as shown in Table II. Then after the pre-processing stage, the next stage is to do the splitting process with the proportion of 70% train data and 30% data test.

TABLE I
DATASET

DATE_STAMP	ASSET_ID	PRESSURE
2020-08-01 01:00:00.000	133071	590.056824
2020-08-01 01:00:00.000	133004	1478.233398
2020-08-01 01:00:00.000	133060	1269.566406
2020-08-01 01:00:00.000	133002	1470.804688
2020-08-01 01:00:00.000	133003	1420.271729

TABLE II
DATASET AFTER PRERPOCESSING

DATE_STAMP	ASSET_ID	PRESSURE
2020-08-01 01:00:00.000	133060	1269.566406
2020-08-01 02:00:00.000	133060	1276.3817138671875
2020-08-01 03:00:00.000	133060	1281.833984375
2020-08-01 04:00:00.000	133060	1290.2850341796875

B. Test Result

1) Model Evaluation

Testing was carried out with the KNN and LSTM models; the dataset from the results of operational records of oil and gas company gas pipeline transmissions located in the Natuna Sea used the "PRESSURE" column or pressure. At the testing stage, the number of random combinations that were tried was 10, and in Table III are the Hyper Parameters used in each model that was built.

TABLE III
HYPER PARAMETERS USED

Random Search Hyper Parameters	
LSTM	**KNN**
Number of random combinations to try = 10	Number of random combinations to try = 10

lstm_units_list = [32, 64, 128]
lstm_dropouts_list = [0.1, 0.2, 0.3]
dense_units_list = [20, 30, 40]
dense_activations_list = ['relu', 'tanh', 'sigmoid']
	n_neighbors_list = [3, 5, 7, 10, 12]
	weights_list = ['uniform', 'distance']
	algorithm_list = ['auto', 'ball_tree', 'kd_tree', 'brute']

When the test dataset is entered into the KNN and LSTM models through the Random Search Hyper Parameter, in Table IV, the KNN model with evaluation metrics R^2 Score 99.46, RMSE 7.543859, MAE 5.385363, and MAPE 4.661628e-03 is obtained from the Random Search Hyper Parameter results with Neighbor = 3, Weight = "distance," and Algorithm = "auto," while in Table V, the LSTM model with evaluation metrics R^2 Score 99.77, RMSE 0.004242, MAE 0.003709, and MAPE 0.004659 obtained from the results of Random Search Hyper Parameters LSTM = 124, Dropout = 0.2, Units = 30, and Actication = "Sigmoid".

TABLE IV
RESULTS OF RANDOM SEARCH HYPER PARAMETERS KNN

Neighbors	Weight	Algorithm	R^2	MAE	MAPE	RMSE
3	**distance**	**auto**	**99.46**	**5.385363**	**4.661628e-03**	**7.543859**
10	distance	brute	99.29	5.489309	4.733994e-03	8.401189
10	uniform	brute	99.22	5.616704	4.859758e-03	8.792731
12	distance	ball_tree	98.99	5.956501	5.184650e-03	10.013267
3	uniform	kd_tree	98.82	5.652836	4.868172e-03	10.743652
10	distance	kd_tree	95.34	5.897086	4.030198e+07	22.034641
3	distance	brute	95.27	5.939401	4.056469e+07	22.191273
7	distance	auto	95.21	5.755424	4.044737e+07	22.214679

7	uniform	kd_tree	95.19	6.052267	3.999832e+07	22.106424
7	distance	kd_tree	95.12	5.656472	4.018136e+07	21.786993

TABLE V
RESULTS OF RANDOM SEARCH HYPER PARAMETERS LSTM

LSTM	Dropout	Units	Activation	R^2	MAE	MAPE	RMSE
128	**0.2**	**30**	**Sigmoid**	**99.77**	**0.003709**	**0.004659**	**0.004242**
128	0.2	20	Sigmoid	99.40	0.004193	0.005648	0.006890
64	0.1	40	Sigmoid	99.40	0.004822	0.006291	0.006872
32	0.1	30	tanh	99.23	0.006799	0.008400	0.007792
32	0.3	20	tanh	99.15	0.005765	0.007700	0.008181
128	0.2	40	relu	99.02	0.003653	0.005339	0.008802
64	0.3	20	relu	98.93	0.005596	0.007624	0.009177
128	0.1	30	relu	96.32	0.004296	0.006999	0.017036
32	0.3	40	relu	96.18	0.008241	0.011698	0.017355
128	0.1	20	relu	95.21	0.009128	0.012588	0.019441

2) *Forecasting*

Then predictions are made on each KNN and LSTM model's training model. Forecasting is done by entering the prepared data train with approximately 6000 data trains. Furthermore, it can be seen in Figure 2 and Figure 3, which is a comparison between the Training Data and the Pressure Prediction Data from the KNN and LSTM models.

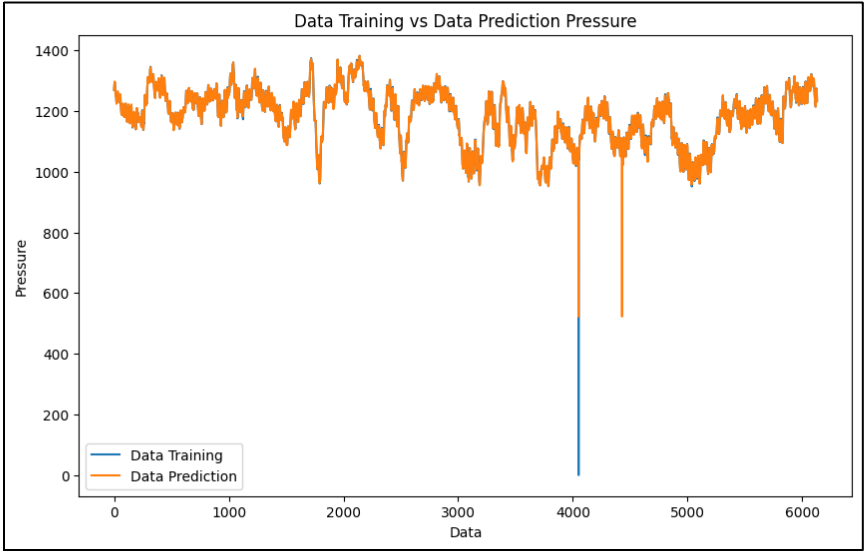

Fig. 2. Comparison of Training Data with KNN Model Prediction Data

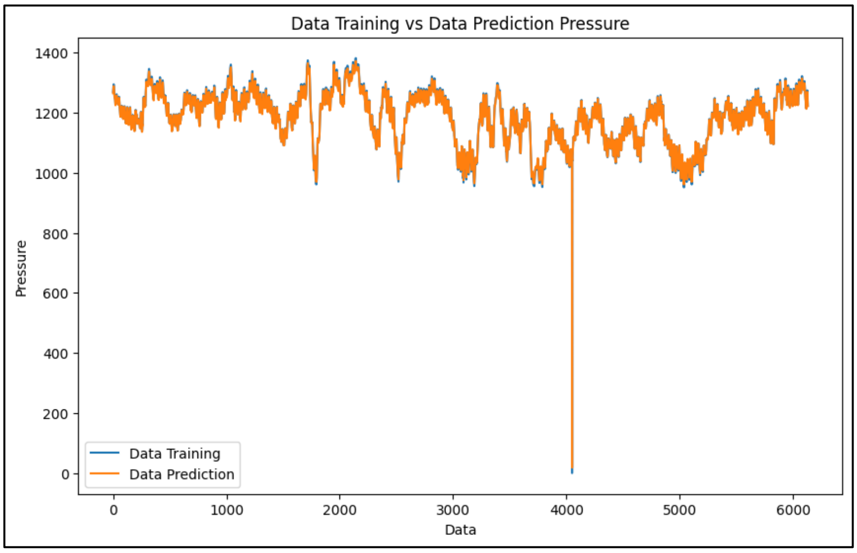

Fig. 3. Comparison of Training Data with LSTM Model Prediction Data

Then predictions are made on each of the best KNN and LSTM models. Forecasting is done by entering test data prepared with more or less test data used as much as 2500 data. Furthermore, it can be seen in Figure 4 and Figure 5, which is a comparison between the Test Data and the Pressure Prediction Data from the KNN and LSTM models.

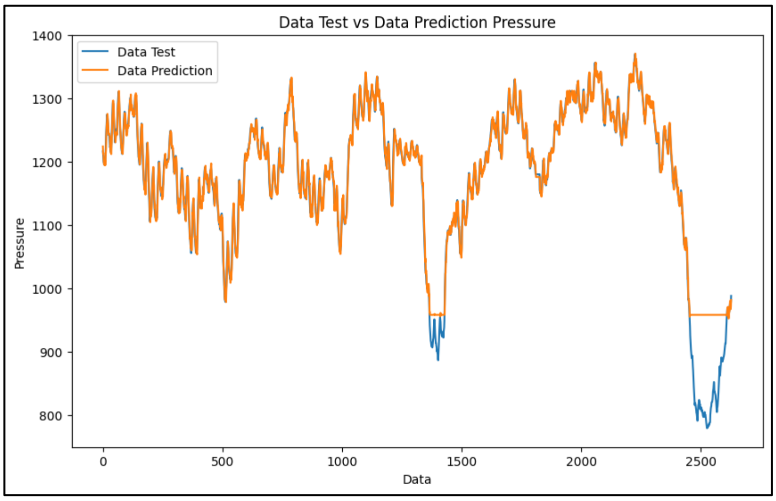

Fig. 4. Comparison of Test Data with KNN Model Prediction Data

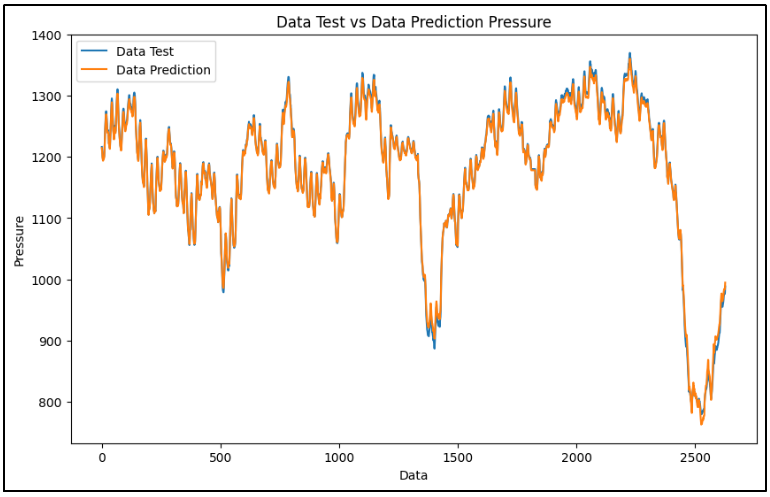

Fig. 5. Comparison of Test Data with LSTM Model Prediction Data

3) *Analysis of Test Result*

Based on Table IV and Table V, it can be seen that the R^2 Score for the LSTM model obtained from the best results of the Random Search Hyper Parameter LSTM = 124, Dropout = 0.2, Units = 30, and Actication = "Sigmoid" is higher than the KNN model obtained from the best result of Random Search Hyper Parameter with Neighbor = 3, Weight = "distance", and Algorithm = "auto". The LSTM model with evaluation metrics RMSE 0.004242, MAE 0.003709, and MAPE 0.004659 is lower than the KNN model with RMSE metric 7.543859. MAE 5.385363, and MAPE 4.661628e-03.

Then in Table VI, the results of the metric evaluation at the KNN and LSTM forecasting model stages can be seen. The forecasting results show that the R^2 score for the LSTM model is higher than that for the KNN model. The LSTM model with evaluation metric RMSE 0.006569, MAE 0.004261, and MAPE 0.005806 is lower than the KNN model with RMSE metric 0.014335. MAE 13.563311, and MAPE 0.014335. So that the LSTM model is more reliable for predicting data in case studies of operational pressure of natural gas pipeline transmission networks.

TABLE VI
EVALUATION RESULTS ON THE KNN AND LSTM FORECASTING MODELS

Forecasting Model	Metrics Evaluation			
	R^2 Score	MAE	MAPE	RMSE
KNN	92.82	13.563311	0.014335	0.014335
LSTM	**99.45**	**0.004261**	**0.005806**	**0.006569**

In [3], evaluating the performance resulting from applying KNN to complete forecasting time series data by considering patterns of data similarity and with the ***k*** value used produces forecast values with different errors in each forecasting period. Whereas [4] evaluates the performance of the implementation of the LSTM model based on the two experimental parameters of the scenario results to get the best LSTM model. The two studies assess the performance of models built using various evaluation metrics.

TABLE VII
PRESSURE PREDICTION RESULTS ON KNN AND LSTM FORECASTING MODELS

Forecasting Model	**Monthly Pressure**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KNN	978,32	978,32	978,32	978,32	978,32	978,32	978,32	978,32
LSTM	982,38	1.102,9	1.175,9	1.226,5	1.264,6	1.294,7	1.319,5	1.340,4

Table VII shows the prediction results from the KNN and LSTM models. The KNN model tends to produce the same value for eight months, while the LSTM model produces pressure values that tend to vary. The results shown in Table VII show that LSTM tends to vary because the LSTM model uses a sequential algorithm. LSTM can remember sequential data patterns that have been studied or used according to the working principle of the LSTM model itself. The working principle enables the LSTM to address problems involving data sequences, such as predicting time series. Meanwhile, the KNN (K-Nearest Neighbors) model has characteristics based on x and y values that allow searching for the closest value from existing data. KNN finds k-neighbors (closest neighbors) of new data based on Euclidean distance or other metrics. Although KNN can provide good results in some cases, these models tend to have lower performance when dealing with high-dimensional data or when the amount of data becomes very large.

IV. Conclusion and Future Works

Based on the test results and analysis of the tests that have been completed, it was found that the LSTM model is better at predicting pressure conditions in natural gas pipeline transmission networks with an, the prediction results from the KNN and LSTM models for the KNN model tend to produce the same pressure value for eight months, while the LSTM model produces pressure values that tend to vary. The previous results show that LSTM outperforms KNN regarding the best model and best prediction for all evaluated error metrics. The hyperparameters and algorithms used in the KNN and LSTM models significantly affect the performance of the two models. The two models that were built obtained an R^2 score above 99%. This finding highlights the advantages of LSTM in predicting data with high dimensions or when the amount of data becomes very large and produces much lower error metrics than KNN. Future studies of the developed model are expected to be able to detect pipe leaks, detect anomalies in operational data, and also determine the actual pressure in pipeline transmission using data from sources. Besides that, it can take an approach from model development such as single-step forecasting, multi-step forecasting, or multi-output.

Acknowledgment

This research was carried out by the Faculty of Informatics, Telkom University.

References

[1] "Pipeline Basics & Specifics About Natural Gas Pipelines," May 2015. Accessed: May 29, 2023. [Online]. Available: https://pstrust.org/wp-content/uploads/2015/09/2015-PST-Briefing-Paper-02-NatGasBasics.pdf

[2] A. Mohd, S. Khan, S. Al, and I. Quraishy, "Pressure Drop Analysis of Natural Gas Transmission Line in Bangladesh."

[3] A. Nadwah and A. Faisol, "IMPLEMENTASI WEIGHTED K-NEAREST NEIGHBOR UNTUK PERAMALAN DATA DERET WAKTU," *J. Ris. & Ap. Mat*, vol. 05, no. 02, pp. 111–117, 2021.

[4] L. Wiranda and M. Sadikin, "PENERAPAN LONG SHORT TERM MEMORY PADA DATA TIME SERIES UNTUK MEMPREDIKSI PENJUALAN PRODUK PT. METISKA FARMA."

[5] C. C. Kurniawan, S. Rostianingsih, and L. W. Santoso, "Penerapan Metode KNN-Regresi dan Multiplicative Decomposition untuk Prediksi Data Penjualan pada Supermarket X."

[6] F. Martínez, M. P. Frías, F. Charte, and A. J. Rivera, "Time Series Forecasting with KNN in R: the tsfknn Package."

[7] I. Arimi, R. Purwaningsih, and Z. F. Rosyada, "METODE K-NEAREST NEIGHBOR UNTUK MEMPREDIKSI PENJUALAN PRODUK PADA UMKM PENGOLAHAN IKAN MAJU JAYA."

[8] Z. Wang, S. Ji, and B. Yu, "Short-Term Traffic Volume Forecasting with Asymmetric Loss Based on Enhanced KNN Method," *Math Probl Eng*, vol. 2019, 2019, doi: 10.1155/2019/4589437.

[9] A. F. Ihsam, Darmadi, S. Uttunggadewa, S. D. Rahmawati, I. Giovanni, and S. N. Himawan, "Multi-Layer LSTM Implementation in Operational Condition Forecasting of a Natural Gas Transmission Pipeline Network," in *ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 244–249. doi: 10.1109/ICOIACT55506.2022.9971837.

[10] S. R. Patra, H. J. Chu, and Tatas, "Regional groundwater sequential forecasting using global and local LSTM models," *J Hydrol Reg Stud*, vol. 47, Jun. 2023, doi: 10.1016/j.ejrh.2023.101442.

[11] F. Bayram, P. Aupke, B. S. Ahmed, A. Kassler, A. Theocharis, and J. Forsman, “DA-LSTM: A dynamic drift-adaptive learning framework for interval load forecasting with LSTM networks,” *Eng Appl Artif Intell*, vol. 123, Aug. 2023, doi: 10.1016/j.engappai.2023.106480.

[12] A. F. Ihsan and W. Astuti, “Deep Learning Based Anomaly Detection on Natural Gas Pipeline Operational Data,” in *2022 2nd International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 228–233. doi: 10.1109/ICICyTA57421.2022.10037988.

[13] A. F. Ihsan, Darmadi, and I. Fahmi, “Analysis of Deep Learning Performance in Hydrocarbon Dew Point Temperature Estimation in Gas Pipeline Network,” in *2022 10th International Conference on Information and Communication Technology, ICoICT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 200–204. doi: 10.1109/ICoICT55009.2022.9914860.

[14] N. R. Sari, W. F. Mahmudy, A. P. Wibawa, and E. Sonalitha, “Enabling external factors for inflation rate forecasting using fuzzy neural system,” *International Journal of Electrical and Computer Engineering*, vol. 7, no. 5, pp. 2746–2756, Oct. 2017, doi: 10.11591/ijece.v7i5.pp2746-2756.

[15] M. N. Akhter *et al.*, “An Hour-Ahead PV Power Forecasting Method Based on an RNN-LSTM Model for Three Different PV Plants,” *Energies (Basel)*, vol. 15, no. 6, Mar. 2022, doi: 10.3390/en15062243.

[16] M. Jamei *et al.*, “Designing a decomposition-based multi-phase pre-processing strategy coupled with EDBi-LSTM deep learning approach for sediment load forecasting,” *Ecol Indic*, vol. 153, Sep. 2023, doi: 10.1016/j.ecolind.2023.110478.

[17] C. Erden, “Genetic algorithm-based hyperparameter optimization of deep learning models for PM2.5 time-series prediction,” *International Journal of Environmental Science and Technology*, vol. 20, no. 3, pp. 2959–2982, Mar. 2023, doi: 10.1007/s13762-023-04763-6.

[18] F. Li and G. Jin, “Research on power energy load forecasting method based on KNN,” *International Journal of Ambient Energy*, vol. 43, no. 1, pp. 946–951, 2022, doi: 10.1080/01430750.2019.1682041.

[19] M. Liu *et al.*, “The applicability of lstm-knn model for real-time flood forecasting in different climate zones in China,” *Water (Switzerland)*, vol. 12, no. 2, Feb. 2020, doi: 10.3390/w12020440.

[20] H. Chen, M. Zhu, X. Hu, J. Wang, Y. Sun, and J. Yang, “Research on short-term load forecasting of new-type power system based on GCN-LSTM considering multiple influencing factors,” *Energy Reports*, vol. 9, pp. 1022–1031, Oct. 2023, doi: 10.1016/j.egyr.2023.05.048.

[21] H. Kusdarwati and S. Handoyo, “System for prediction of non stationary time series based on the wavelet radial bases function neural network model,” *International Journal of Electrical and Computer Engineering*, vol. 8, no. 4, pp. 2327–2337, Aug. 2018, doi: 10.11591/ijece.v8i4.pp2327-2337.

[22] Y. Abdillah and S. Suharjito, “Failure prediction of e-banking application system using adaptive neuro fuzzy inference system (ANFIS),” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, p. 667, Feb. 2019, doi: 10.11591/ijece.v9i1.pp667-675.