Deteksi Anomali Menggunakan *Deep Learning* Berbasis LSTM pada Data Operasional Pipa Gas Alam

Muhammad Rieza Fachrezi[1], Dr. Aditya Firman Ihsan, S.Si., M.Si.[2], Widi Astuti, S.T., M.Kom.[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]riezacornelis@students.telkomuniversity.ac.id, [2]adityaihsan@telkomuniversity.ac.id,
[3]widiwdu@telkomuniversity.ac.id

Abstrak

Industri minyak dan gas alam merupakan sektor yang krusial dalam kehidupan sehari-hari. Insiden yang tidak diinginkan dalam sektor ini dapat berdampak signifikan pada sektor rumah tangga. Oleh karena itu, sistem peringatan dini otomatis diperlukan untuk mendeteksi kesalahan dalam jaringan pipa yang menghubungkan titik produksi dan pemrosesan. Metode deteksi anomali dapat digunakan untuk mengatasi masalah ini. Salah satu model yang cocok untuk deteksi anomali *unsupervised* adalah model *Long Short-Term Memory* (LSTM) berbasis *deep learning*. Penelitian ini bertujuan untuk mengimplementasikan model deteksi anomali berbasis LSTM pada data operasional *time-series* dari titik pengamatan yang terletak di dalam hilir, yang terdiri dari 17 fitur yang dapat digunakan dan 8.736 baris data, mewakili data selama satu tahun. Pemilihan model melibatkan pengoptimalan *hyperparameter* (misalnya *dropout, regularizer, layer,* dan *batch size*) menggunakan *Mean Squared Error* (MSE) melalui *3-fold cross validation,* menghasilkan 10 kandidat model. Model dengan performa terbaik kemudian dilatih menggunakan data pelatihan. Setelah pelatihan, model LSTM merekonstruksi data *time-series* asli untuk menghitung skor anomali berdasarkan metrik jarak *Euclidean*. Skor ini menentukan anomali menggunakan ambang batas yang ditetapkan dari distribusi skor anomali. Interpretasi manusia memvalidasi kemampuan model untuk secara akurat mengidentifikasi anomali dalam dataset. Persyaratan infrastruktur untuk penerapan di dunia nyata juga dibahas, dengan fokus pada penggunaan metode *edge computing* untuk meningkatkan kemampuan deteksi anomali secara *real-time*.

Kata kunci : deteksi anomali, LSTM, minyak dan gas bumi

Abstract

The oil and natural gas industry is a crucial sector in our daily lives. Unwanted incidents in this sector can significantly impact the household sector. Therefore, an automatic early warning system is necessary to detect errors in the pipeline network connecting production and processing points. Anomaly detection methods can be used to overcome these problems. One model suitable for unsupervised anomaly detection is the deep learning Long Short-Term Memory (LSTM) model. This research aims to implement an LSTM-based deep learning anomaly detection model on time-series operational data from an observation point situated within the sink, which consists of 17 usable features and 8,736 data points, representing a year's worth of data. Model selection involves optimizing hyperparameters (e.g., dropouts, regularizers, layers, and batch sizes) using the Mean Squared Error (MSE) through 3-fold cross-validation, resulting in 10 model candidates. The best-performing model is then trained using the training data. After training, the LSTM model reconstructs the original time-series data to calculate anomaly scores based on the Euclidean distance metric. These scores determine anomalies using a set threshold derived from the distribution of anomaly scores. Human interpretation validates the model's capability to accurately identify anomalies within the dataset. Infrastructure requirements for real-world applications are also discussed, focusing on the use of edge computing methods to enhance real-time anomaly detection capabilities.

Keywords: anomaly detection, LSTM, oil and natural gas

1. Introduction

The oil and natural gas industry is an essential industry since oil and gas is an energy intermediary used by the industrial sector all the way to the household sector [1]. The distribution of oil and gas must therefore be efficient and accurately measured using measuring devices so that supervision and evaluation can be carried out if harmful things occur because the impact can be felt by almost everyone.

The process of oil and gas distribution requires pipelines that connect from upstream to downstream. Of course, the installed pipes have the possibility of unwanted things happening, such as pipe leaks, failure of measuring instruments, or measuring sensor errors. Therefore, a reliable system is needed that can be utilized as an automatic early warning in case of such events.

One of the methods that can be used as an early warning is the machine learning method of anomaly detection. Anomaly detection, especially for time-series, is the detection of unexpected system behavior as the system runs [2]. Anomaly detection can be done supervised, unsupervised, and semi-supervised [3]. The problem with anomaly detection modeling is that the training data is often unbalanced and very little data has an anomaly label. Therefore, most anomaly detection models must be trained unsupervised [2].

Unsupervised machine learning does not require input labels, unlike supervised machine learning which requires input labels for classification. As such, unsupervised anomaly detection methods automatically flag data that are considered abnormal and do not need to be labeled from the beginning indicating whether the data are anomalous or not [1][3].

In the domain of anomaly detection, there are a variety of systems that have been successfully implemented on time-series datasets. These systems range from statistical methods, classical machine learning, to deep learning. Examples of statistical type methods include ARMA [4] and ARIMA [5]. The classical machine learning type includes PCA [6], k-Means [7], and K-nearest neighbor (KNN) [8].

The most recent type of approaches are deep leaning methods which include autoencoder (AE) [1], variational autoencoder (VAE) [9], and long short-term memory (LSTM) [10]. There are also anomaly detection systems that combine several methods, such as Hybrid KNN [11], VAE-LSTM [2] which obtained 100% recall on all tested datasets, and a combination of VAE and Generative Adversarial Network (GAN), namely VAE-GAN based on LSTM [12].

Combining machine learning techniques can significantly outperform individual methods. Lin et al. [2] demonstrated this by showing that their hybrid VAE-LSTM model achieved superior anomaly detection performance compared to standalone VAE, LSTM, or ARMA models. Notably, the VAE-LSTM achieved a remarkably high recall of 100% across all tested datasets, suggesting its effectiveness in identifying anomalies. These findings suggest that combining models with complementary strengths can be beneficial for anomaly detection tasks.

In this paper, we implement anomaly detection system using deep learning method based on LSTM using operational data of oil and natural gas pipelines. This choice is motivated by LSTM's well-established capability to capture long-term dependencies within sequential data [2], a crucial aspect for analyzing time series data like operational data from oil and natural gas pipelines. Then, we analyze the performance of the model using hyperparameter tuning. In the end, we evaluate the results by using human interpretation as to whether the data points are real anomalies or not.

2. Methodology

2.1 Dataset

This study utilizes pipeline operational data from an observation point situated within the sink, acquired secondarily from an oil and natural gas company. Measuring instruments installed at the observation point continuously record data streams reflecting the physical state (pressure, temperature) and composition (ethane, propane, etc.) of the gas. This data includes features like gas pressure, temperature, energy rate, and volume rate, along with compositional data for various hydrocarbons including ethane, propane, isobutane, butane, isopentane, pentane, and heavier alkanes like hexane, heptane, octane, and nonane. Additionally, data for nitrogen, carbon dioxide, and water are included.

The dataset was originally recorded at one-minute intervals. To facilitate computational efficiency and minimize the influence of short-term variations, the data was down sampled to hourly entries. This yielded a more manageable dataset encompassing approximately 8,736 data points, representing a year's worth of information. Due to the wide spread of values across different features, the data was further scaled to a [0, 1] range using min-max scaling. This ensures all features contribute equally during model training and avoids biases caused by features with significantly larger or smaller values.

A subset of 17 features will be used for training. Two features, methane (C1) and hydrocarbon dew point (HCDP), were excluded due to having exclusively missing data across all the data points used in this subset. However, it's worth noting that these features might have been recorded in other observation points. The data will be divided for training, validation, and testing. The majority (53%) will be used as training data to fit the machine learning model. The remaining data will be split into a 13% validation set and a 33% testing set. The validation set acts as a control to monitor for overfitting during training, while the testing set provides a final assessment of the model's ability to detect anomalies on unseen data.