1. Introduction

In the energy sector, natural gas is an important and very efficient energy source for many aspects of daily life. In addition to supplying domestic energy needs, the industrial and power production sectors also depend heavily on the widespread usage of natural gas. The advantage of natural gas being considerably more environmentally friendly compared to other fossil fuels makes it a top and preferred choice in efforts to reduce carbon emissions. Therefore, monitoring in the operational of this industry is of utmost importance.

The data generated from various sensors and monitoring systems in the gas operational pipeline is large and complex. Every single second, these sensors record different types of data, ranging from operational condition such as pressure and temperature readings to gas flow rates and chemical composition. This data is then analyzed to ensure operations run smoothly and safely. Although the fact that monitoring is carried out continuously, there are many risks than can still occur, such as gas leaks, sudden pressure drops, or equipment failure. Because of that, a dependable system is required to identify abnormalities or odd events within the operating data. This is important because these kinds of interruptions can have detrimental effects on the environment, endanger the safety of workers, and result in large financial losses.

Methods for anomaly detection have been widely discussed in the scientific literature, but their implementation in some industries, including the natural gas sector is still relatively limited. These methods can be divided into a number of groups. Some of these methods are statistical methods such as Principal Component Analysis (PCA) which is used to reduce the dimensionality of data and identify anomalous patterns based on data variability [1]. Other methods, like Support Vector Machine (SVM) [2] and Knearest Neighbor (KNN) [3] can be grouped as Classificationbased method. These methods use a supervised learning approaches to classify data into normal or anomaly data point. Meanwhile, clustering-based methods like K-Means [4] and DBSCAN [5] group data points into clusters and detect anomalies as data that does not fit into any cluster. Other than these traditional methods, the use of deep learning methods approaches to detect anomaly has been increasing. Some of the approaches, like autoencoders and recurrent neural network (RNN), have proven to be able to process large amounts of data and identifying intricate anomaly patterns with high accuracy [6].

Currently, there are few studies for anomaly detection on natural gas pipeline operational data that have been conducted. Reference [7] implements several methods for detecting anomaly in this sector. The methods used in this research include Random Forest, SVM, KNN, Gradient Boosting, and Decision Tree.

While currently there are a lot of approaches that have been developed for anomaly detection, a research by Schmidl et al., shown that there is no one-size-fits-all approach that can consistently performs better than other methods in every circumstance [8]. Every approach has pros and cons that vary depending on the data properties and the particular environment in which it is used. Even old established techniques such as PCA and K-Means used by many and can compete under certain circumstances with more recent methods. This suggests the importance of carefully considering various factors, ranging from variety of specific factors such as data type, real-time requirements, and computational complexity when selecting an appropriate anomaly detection method.

The Extended Isolation Forest method is the anomaly detection strategy we have used for this study [9]. Reference [10] used this method to detect anomaly in small hydroelectric plants data, while [11] used this method to detect anomaly for artificial pancreas system. This method offers an improvement over the original Isolation Forest algorithm [12], which became well-known for its effectiveness in detecting anomaly in huge and complex datasets. The Extended Isolation Forest expands upon the basic foundational framework of Isolation Forest by incorporating a branching mechanism that involves random rotation, which makes the method capable of handling more varied types of data, including multidimensional timeseries data such as natural gas operational data. The algorithm works by building a decision tree that separates normal and anomalous data based on how easily it can be isolated within the tree structure. Because of this ability, the Extended Isolation Forest is expected to detect anomalies more accurately and efficiently.