1. INTRODUCTION

Oil and gas irreplaceability as a source of energy has come to be shown by the continued progress of the oil and gas industries. sectors such as the transportation sector and the household sector needed oil and gas. Pipelines are used as a means of distribution to distribute oil and gas, where the quality and reliability of the pipeline are important factors that determine the success of the oil and gas industry. Therefore, maintaining the stability of these pipelines through monitoring is required to prevent the occurrence of undesirable events.

Pipeline damage can cause considerable risks, which include threats to human health, workplace safety, and the environment, including issues such as ozone depletion and climate change [1]. Furthermore, these damages can cause considerable financial setbacks for the oil and gas sector [2]. Several factors, such as rust, corrosion, and leakage, contribute to pipeline damage. For this reason, timely identification of anomalies is necessary for the early detection of potential problems in oil and gas pipeline distribution. However, manual detection methods are time consuming. A viable solution to this task is to create a machine learning model for anomaly detection, which facilitates early identification of potential problems during the distribution process.

One of the machine learning methods for anomaly detection in the oil and gas industry pipeline is through unsupervised learning. The approach works by analyzing the intrinsic pattern in the data [3]. By using this approach, we can identify the normal pattern in the data, and then detect the anomalous data points. However, the detected anomalies do not necessarily mean a significant risk to the pipeline, as they can originate from sensor errors or poor streaming data. Therefore, collaboration with experts in the field is needed to support real-time anomaly detection and provide accurate interpretation of the detection results.

Alongside unsupervised learning, supervised learning can also be implemented for anomaly detection in oil and gas pipelines. This approach requires labeled data where anomalies are explicitly marked. The methods that use this approach are support vector machine (SVM) [4] and K-nearest neighbor (KNN) [5]. However, such methods can be challenging as they require a deeper understanding of a particular domain to be able to accurately label the data, as they rely on the knowledge and features of anomalies present in said domain. Other traditional methods with modifications, like one class support vector machine (OCSVM) [6] and isolation forest [7] can also be used to detect anomalies.

Deep learning models are popular in the current state of anomaly detection. The model usually reconstructs the input data using a deep representation of the neural network. The anomalies are marked with high reconstruction errors. The commonly used deep learning architecture to reconstruct said data is by using an autoencoder [8]. In addition to autoencoders, other deep learning architectures, such as RNN [9], can also be adapted for anomaly detection tasks. Basic RNN models have the ability to capture temporal dependencies, but they suffer from vanishing gradient problems [10] that reduce their capability. That problem poses a limitation for basic RNN in retaining information from earlier states. To deal with that problem, more advanced RNN architectures have been developed, namely LSTM [11] and Gated Recurrent Unit (GRU) [12].

That being said, according to Schmidl et al.'s experiments, there is no dominant model that outperforms all the tested data out of the 71 available algorithms. [13]. Even with the increasing popularity of deep learning in anomaly detection, there aren’t many that apply it to the oil and gas industry. Traditional techniques like support vector machines (SVM), random forests, gradient boosting, and closest neighbor are often used. Aljameel et al has used and compared it in their experimentation [1].

This paper will implement an autoencoder detection model using RNN and GRU for natural gas and oil pipeline operational data. These anomalies could be present in one or multiple linked attributes. By analyzing the performance of each model and tuning its parameters, we aim to minimize the reconstruction error and compare the two models to decide which one is better in terms of performance. After that, we can finally evaluate the result with human interpretation to determine if the datapoint is a real anomaly or not.