EPiC Series in Computing

Volume 79, 2021, Pages 21–29

Proceedings of ISCA 34th International Conference on Computer Applications in Industry and Engineering

EPiC Computing

Generative Adversarial Neural Network and Genetic Algorithms To Predict Oil and Gas Pipeline Defect Lengths

Huda Aldosari[1], Sanguthevar Rajasekaran[2] and Reda Ammar[3]
[1,2,3] University of Connecticut
[1,2,3]{Huda.aldosari,sanguthevar.rajasekaran,reda.ammar@uconn.edu}

Abstract

Estimation of expected failure in an oil and gas pipeline system is challenging due to large uncertainties in the parameters associated with burst failure predictive models. The development of machine learning (ML) algorithms for reliability and risk assessment applications has attracted considerable attention from the scientific and research community in recent years. Working on the automation, efficiency, and optimization of underground oil and gas pipeline networks demands open access to extensive databases, which may not be possible. Oil and gas databases are confidential assets of specific countries, and no one can access these databases easily. As a result, training ML models is a big challenge, since it needs large data. To address this data shortage, in this paper, we have generated synthetic training datasets using a tabular generative adversarial neural network (TGAN). The generated synthetic data and real data (when available) were combined to train an artificial neural network (ANN). To further enhance the performance of the proposed system, the application of a genetic algorithm (GA) has been introduced to optimize the weights and biases of the ANN automatically. The results show superior performance results when compared with the previously reported algorithms in the literature. The proposed methodology succeeds to predict Oil and Gas pipeline defects with robust results and low error rates.

1 Introduction

Transportation of gas and oil through a proper pipeline system is considered the safest, reliable, and economical method. That is why it is indispensable to maintain the integrity of pipelines to avoid any catastrophic incident. However, aging is the crucial factor for high-frequency pipeline failures in recent years [1]. This is because the corrosion deterioration reduces the strength of pipelines. Thus, corrosion has been categorized as the fundamental threat to the safety and performance of oil and gas pipelines [2]. Reduction in material strength due to excessive corrosion growth may result in burst system failure causing substantial social, environmental, and economic consequences. Burst failure can be referred to as the ultimate pressure of loss that a pipe can tolerate before it collapses. In other words, it is the maximum load-bearing capacity of a line under internal pressure. In the various past, systems have been constructed to detect burst failures due to corrosion. These systems mostly use the concept of semi-empirical fracture mechanics in which losses are dependent upon the flow stress. Thus, failures can be detected by comparing the maximum fluid pressure and actual failure pressure inside the pipe [3].

Various uncertainties are involved in the estimation of pipeline failure because of the different associated parameters with burst failure predictive models [4]. Oil and gas pipelines are composed of hundreds of segments, making them typically complex and large in structure, making the physical probabilistic approach a computationally intensive process [5]. A detailed reliability analysis is not feasible in difficult gas and oil pipeline systems to decide on maintenance. A considerable effort is being put by the utilities for developing an efficient, cost-effective, safe, and reliable method for the proper maintenance of pipeline networks as the available approaches for risk and failure detection in pipelines are not efficient computationally. That is why a robust and efficient technique is needed to timely predict the defect and pipeline failure as a replacement for physical analysis [6]. Data-driven machine learning (ML) algorithms provide a viable alternative computational approach for risk and failure analysis of gas and oil pipeline networks [7,8]. ML can be defined as a system based on a semi-automated process capable of observing the data for developing an algorithm. Such ML-based approaches have been applied for risk analysis purposes in various engineering applications [9-13]. Algorithms like K-nearest neighbor, support vector machine, and random forest have emerged as widely acceptable algorithms in the last decade [14].

Anghel et al. (2012) developed a novel support vector machine to predict the failure probabilities in pipeline networks with corrosion defects [15]. He developed a procedure based on classification reliability and established a link between reliability methods and artificial intelligence. This procedure provided a simple but viable alternative approach to prioritize the maintenance actions. Burton and Haung (2020) applied ML techniques to classify the failure modes of concrete-based frames with infill walls [16]. Winkler et al. (2008) predicted the failures in water pipes using decision trees [17]. Boosting and bootstrap aggregation techniques are applied to enhance the model's accuracy and measure the model confusion matrix's performance. Mishra et al. (2020) utilized different ML algorithms for identifying the modes of failure in concrete-based bridge columns [18]. The efficacy of 6 different ML algorithms was compared by the author using an experimental dataset. ANN proved to be the most efficient approach among the six algorithms. Kiani et al. (2019) explored ML models for the derivation of fragility curves [19]. Siam et al. (2019) used clustering algorithms with a limited number of datasets to classify masonry walls [20].

Data-driven techniques were applied by Mangalathhu et al. (2020) for recognizing the mode of failure in concrete shear walls [21]. Ten input and output parameters are utilized to capture the material, reinforcement, and geometric characteristics to identify failure modes. Global accuracy, recall, and precision are some of the standard metrics used to evaluate the efficiency of ML algorithms. The study presented in [21] reports promising results for the classification of failure modes that can be applied to other infrastructures. Other ML algorithms by Qi and Zhu et al., Jeon et al., and Zhang et al. have the potential to solve complex problems without any demand on the mechanical analysis [22-25]. Timely prediction of the expected failures in pipeline structures can help the management team and decision-makers in taking necessary intervention measures to avoid the worst consequences of burst failure.

In this paper, we employ ML models and the genetic algorithm to solve the problem of predicting failures in pipeline systems. There are three main goals of this study: (1) To create a comprehensive dataset within a realistic combination and range; (2) To use the generated dataset for assessing the feasibility of ANN to evaluate the failure risks of a pipeline; and (3) To optimize the parameters of an ANN using the genetic algorithm.

2 Methodology

2.1 Dataset

Nondestructive Inspection (NDE) methods are frequently used to screen pipelines for potential defects. A magnetic sensor attached to a computing unit is one such NDE instrument routinely sent for testing oil and gas supply tubing. Magnetic sensors, spaced every 3 mm along the pipeline's circumference, are used to calculate Magnetic Flux Leakage (MFL) signals. The study reported in [26] used a pre-recorded MFL dataset of noiseless and noisy MFL signals at 3dB, 5dB, and 10dB signal-to-noise ratios (SNRs). Since the amount of MFL data is too large, feature extraction techniques have been used to reduce the data's feature space. Using any of the derived functions, on the other hand, does not always result in stronger reliability test outcomes. The most important characteristics are thus selected and fed into the identifying and sizing units.

2.2 Feature Extraction

Feature extraction aims to reduce the size of the MFL data. The MFL signal is expressed in the axial, radial, and tangential dimensions. For each component, statistical and polynomial series are used as feature extraction methods. From original MFL signals, statistical characteristics of the integral normalized signal (INS), mean average (MA), maximum magnitude (MM), standard deviation (STD), and peak-to-peak distance (PPD) are derived, and so are polynomial rank 3, 6, and 6 series corresponding to the axial, radial, and tangential components, respectively. Polynomial coefficients make up the input characteristics, in addition to the five statistical features. There are a total of 33 functions because of this.

2.3 Feature Reduction

Many researchers have shown that different attributes may have differing degrees of discrimination capacity. In some instances, using the whole collection of retained features in the training phase results in poor defect prediction accuracy. Some attributes may hurt the prediction accuracy. As a result, determining the main qualities that make the ML model perfect is a routine practice. Following that, analyses are carried out to assess the suitability of each element for the defect length prediction mission. The most appropriate features that yield the most effective defect convergence rate are then found using neighborhood component analysis (NCA) as an input function pattern for the ML model. The features with a weight close to zero make little difference in calculating defect sizes.

Furthermore, for assigning weights for the retained functions, a principal component analysis (PCA) based weight correlation technique is used. The highest-weighted attributes are then used to train the model. From the axial, radial, and tangential elements, we choose nine characteristics based on their weights: INS and PPD; INS, PPD, and MM; and INS, PPD, MM, and STD from x, y, and z components, respectively.

2.4 Synthetic Data Generation

A Generative Adversarial Network (GAN) has two networks: a generator and a discriminator. The generator learns to generate instances from a true data generating distribution and the discriminator evaluates the instances generated by the generator. They work in a competitive zero-sum game framework.

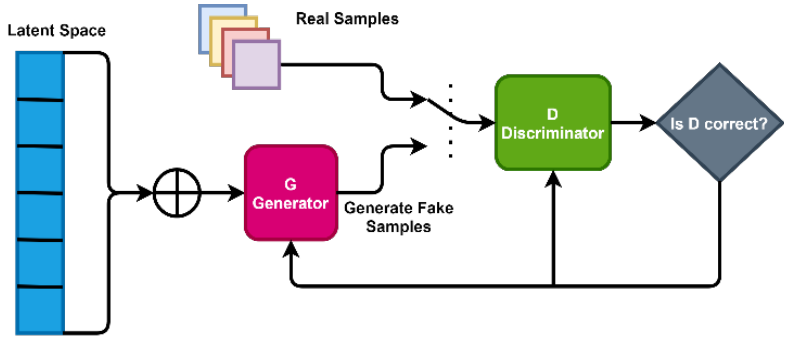

Figure 1: GAN model for the prediction of tabular data

The 'Generator' function is to create the samples that will not be distinguished from the actual examples by 'Discriminator.' During the application of GAN training, speed and data quality in the specific domains are the main problems. These objectives can be achieved through tabular GAN (TGAN) as data background in particular disciplines of ML applications are essential, and GAN fails to generate [27]. The generator model summarizes the distribution of the input data and based on the summarized distribution new data samples are predicted. While the discriminator model predicts the generated data as fake and real and based on the classification the corrections are made.

- **Preprocessing the numerical variables:** Neural network models can generate values with the central distribution over (-1,1). They cluster the numerical variables through training a gaussian mixture model (GMM) with "m" components for each cluster "c." Eventually, GMM is utilized for the normalization of "c." The probability of "c" is computed for each "m" and gaussian distribution as vector "U."
- **Preprocessing the categorical variables:** Because of low cardinality, the probability distribution can be generated using SoftMax. But it is essential to convert the categorical variables into one-hot encoding with noise-binary variables. After the preprocessing of categorical variables, T is converted in required columns to from given vectors. This is an output vector of generator and input for discriminator in GAN.
- **Generator:** Numerical variables are generated in two steps. In the first step, a scalar value V is created, and then the U-cluster vector is formed using the tanh function. Categorical features are developed as the probability distribution over all the possible labels. SoftMax is used for the generation of desired row and an LSTM with the attention mechanism used.
- **Discriminator:** A multi-layer perceptron (MLP) with the Leaky ReLU (LRELU) and batch normalization is used. The first layer has the concatenated vectors with the mini-batch diversity and the feature vector from the LSTM. ADAM is the loss function divergence term for the variables with log loss sum ordinal function.

TGAN is reasonably robust and outperforms Bayesian networks with significantly less average performance between the natural and synthetic data. Using the TGAN model, we generated 50,000 synthetic data samples corresponding to clean, 3dB, 5dB, and 10dB data based on the original dataset. The generated data is then used for further processing along with the original data.

2.5 Optimization by Genetic Algorithm

A Genetic Algorithm (GA) is an evolutionary algorithm. In this algorithm, random changes are applied to the currently available solutions to generate new ones. The concept of GA is based on Darwin's theory of biological evolution. Besides this, GA makes slight alterations to its' solutions, automatically optimizing and getting the best possible solution. GA is applied to the population consisting of specific possible solutions. Population size is equal to the number of solutions. Every solution here is called an individual, and every individual solution contains a chromosome. The chromosome is then represented as the set of parameters called the features defining the individual. The chromosome has a set of genes, and every gene is then represented as a string of the zeros and ones. The genetic optimization approach has been presented to handle the constraints based on the behavioral-memory paradigm. It starts with a random population and sets a constraint counter, and then evolves the population until a feasible flip threshold for the constraint is obtained. Then in the next phase, unsatisfactory constraint points are eliminated from the population. This iteration continues until the objective function is optimized. This GA algorithm demands a linear order of all the constraints. But the effect order of the controls is vague as it has been observed that different declaration of restrictions produces different results.

When the change in the fitness function from one iteration to the next is < 10^{-10}, the algorithm terminates. The objective function of the utilized GA model can be described by the equation (1). Fifty generations with a stall generation limit of 100 were used to optimize a feed-forward ANN using GA. For a feed-forward network with n hidden neurons, 9n+n+1 quantities are required in the weights and biases column vector based on nine features of the dataset.

$$Error = \sum_{j=n} (f_{calculated}(x_j) - f_{model}(x_j)) \quad (1)$$

In this study, GA has been incorporated with a feed-forward artificial neural network (ANN) for regression analysis on the combined dataset (original and synthetic data). Each dataset has 50,000 samples of synthetic data and 1,277 samples from the original dataset. The combined dataset was randomly divided so that 70% of the data was used for training the ANN model, 15% was used for validation, and the remaining 15% was reserved for testing the trained model. The results of the trained model are presented in the results section. The results presented are for the testing sets. For ANN feed-forward neural network with a single hidden layer and 10 hidden neurons, the levenberg-Marquardt training algorithm and log transfer function were used. Python 3.8 was used for GAN training and MATLAB 2020 has been used for the rest of the experiments.

3 Results

In this section, we present the qualitative and quantitative assessment of TGAN, GA, and ANN models. To assess the TGAN, mean absolute error (MAE) between the actual and generated data has been measured corresponding to each feature. Similarly, to evaluate the effectiveness of ANN incorporated with GA, mean square error (MSE) and root mean square error (RMSE) have been measured between actual and predicted data labels.

3.1 GAN network

A tabular generative adversarial network was used to generate synthetic data for MFL based oil and gas pipeline data. Initially, we had 33 features corresponding to the defect length for each dataset as described previously. The original 33 features were further reduced to 9 with the application of PCA, and redundant features were removed from the dataset. The noisy and noiseless reduced datasets with actual defect length were then fed to the TGAN model to generate similar datasets synthetically. TGAN model developed 50,000 samples for each dataset. The MAE between generated and actual samples corresponding to each dataset is then calculated and presented in Table 1.

Features	**Mean Absolute Error (MAE)**			
	Noiseless	**3dB**	**5dB**	**10dB**
1	2.19776	2.616381	3.662933	5.4944
2	1.9178	2.283095	3.196333	4.7945
3	0.76972	0.916333	1.282867	1.9243
4	0.1036	0.123333	0.172667	0.259
5	0.096	0.114286	0.16	0.24
6	0.06264	0.074571	0.1044	0.1566
7	0.04448	0.052952	0.074133	0.1112
8	0.00804	0.009571	0.0134	0.0201
9	0.01212	0.014429	0.0202	0.0303

Table 1: Mean absolute error (MAE) between original & synthetically generated features using TGAN

A higher value of MAE represents a poor performance of the TGAN model and a low correlation between generated and actual data samples. From the table 1, it can be observed that the utilized TGAN model performed significantly well on features 3 to 9 for all datasets by yielding low MAE scores. In comparison, the model performed poorly on features 1 and 2 corresponding to all features. It is also observed that the MAE value of the synthetic data increases as the noise of the signals increases.

3.2 Defect Length Prediction

Defect length was predicted using a feed-forward neural network with a single hidden layer, ten hidden neurons, the Levenberg-Marquardt training algorithm, and log transfer function. Figure 2 depicts the regression plots of all the datasets corresponding to the training, validation, and testing phase. It can be observed that for noiseless and noisy data at 10 dB, the model predicts the defect length perfectly with a good correlation whereas, the model has poor performance for the 3 dB and 5 dB datasets.

The fundamental aim of the study was to optimize the utilized ANN such that the performance of the model increases for all the datasets irrespective of the noise embedded in the MFL signals. For the said purpose, the application of GA has been utilized. After optimizing the weights and biases of the used ANN model, the model's performance increases significantly. Table 2 illustrates the MSE and RMSE of the optimized model after introducing GA.

Lower values of MSE and RMSE indicate better performance of the model on the testing datasets. Table 2 shows that for all datasets, irrespective of the level of noise embedded in the MFL signals. The GA-based optimized ANN model yields good performance results.

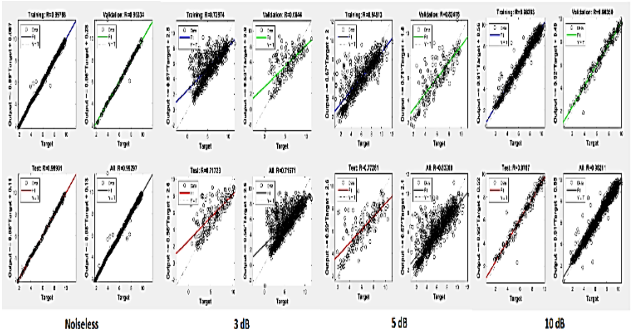

Figure 2: Regression plots depicting the correlation between the actual and predicted features corresponding to all datasets.

	Noiseless	3dB	5dB	10dB
MSE	0.0563	0.5021	0.3391	0.0637
RMSE	0.2374	0.7086	0.5823	0.2525

Table 2: MSE & RMSE offered by trained ANN model after GA optimization.

3.3 Comparison With Previous Studies

To illustrate the significance of utilizing a larger dataset and automated optimization for ANN, the study results have been compared with previously published studies on the same dataset [7,12]. Both studies used a feed-forward neural network with a single hidden layer and 10 hidden neurons, and a levenberg-Marquardt training algorithm. Table 3 summarizes the MSE scores of recently published results on the same datasets and illustrates the significance of the proposed method.

Method	Noiseless	3dB	5dB	10dB
ANN	0.0612	2.8132	1.8175	0.1861
Manual ANN Optimization	0.0139	2.1044	1.1569	0.0947
Proposed Method	**0.0563**	**0.5021**	**0.3391**	**0.0637**

Table 3: Comparison with the previous studies for all the used datasets. The table reports the MSE values for comparison with other datasets.

From Table 3, it can be observed that the proposed method outperforms the previous results by a significant margin on all the datasets except at noiseless data. The most notable feature of the proposed methodology is that it can be used for any kind of MFL data irrespective of the noise level embedded in the signal. Previous methods have shown poor performance results for noisy datasets at 3dB and 5dB, as shown in Table 3. However, the proposed method has low RMSE scores for all the datasets indicating the significance of automated optimization of the ANN model.

4 Conclusions

In this paper, we have developed ANN and GA-based techniques to predict defect lengths for oil and gas pipelines utilizing MFL signals. We have introduced the TGAN model first to increase the training dataset to avoid overfitting and make the proposed model robust. The generated synthetic data was then fed to an ANN to predict the defect lengths. Further, to increase the performance of the model, an application of GA has been incorporated. The proposed methodology yields more stable and robust results for MFL signals embedded in different noise levels and outperforms previously reported results on the same dataset.

References

[1]. Nešić, S., 2007. Key issues related to modelling of internal corrosion of oil and gas pipelines–A review. Corrosion science, 49(12), pp.4308-4338.

[2]. Shahriar, A., Sadiq, R. and Tesfamariam, S., 2012. Risk analysis for oil and gas pipelines: A sustainability assessment approach using fuzzy based bow-tie analysis. Journal of loss prevention in the process Industries, 25(3), pp.505-523.

[3]. Ashraf, H., Waris, A., Gilani, S.O., Tariq, M.U. and Alquhayz, H., 2021. Threshold Parameters Selection for Empirical Mode Decomposition-Based EMG Signal Denoising. INTELLIGENT AUTOMATION AND SOFT COMPUTING, 27(3), pp.799-815.

[4]. Dey, P.K., Ogunlana, S.O. and Naksuksakul, S., 2004. Risk-based maintenance model for offshore oil and gas pipelines: a case study. Journal of Quality in Maintenance Engineering.

[5]. Aldosari, H., Elfouly, R.S., Ammar, R. and Alsulami, M., 2020, March. New Monitoring Architectures for underwater oil/Gas Pipeline using Hyper sensors. In CATA (pp. 307-316).

[6]. Aldosari, H., Elfouly, R., Ammar, R. and Alsulami, M., 2020, July. Performance of New Monitoring Architectures for Underwater Oil/Gas Pipeline using Hyper-Sensors. In 2020 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.

[7]. H. Aldosari, R. Elfouly and R. Ammar, "Optimal Artificial Neural Network Model For Prediction of Oil and Gas Pipelines Defect Length," in 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020 pp. 1457-1462. doi: 10.1109/CSCI51800.2020.00272

[8]. Velázquez, J.C., Caleyo, F., Valor, A. and Hallen, J.M., 2009. Predictive model for pitting corrosion in buried oil and gas pipelines. Corrosion, 65(5), pp.332-342.

[9]. Ashraf, H., Waris, A., Jamil, M., Gilani, S.O., Niazi, I.K., Kamavuako, E.N. and Gilani, S.H.N., 2020. Determination of Optimum Segmentation Schemes for Pattern Recognition-Based Myoelectric Control: A Multi-Dataset Investigation. IEEE Access, 8, pp.90862-90877.

[10]. Asif, A.R., Waris, A., Gilani, S.O., Jamil, M., Ashraf, H., Shafique, M. and Niazi, I.K., 2020. Performance Evaluation of Convolutional Neural Network for Hand Gesture Recognition Using EMG. Sensors, 20(6), p.1642.

[11]. H. Aldosari, R. Elfouly and R. Ammar, "Evaluation of Machine Learning-Based Regression Techniques for Prediction of Oil and Gas Pipelines Defect," in 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020 pp. 1452-1456.doi: 10.1109/CSCI51800.2020.00271

[12]. Ashraf, H., Waris, A., Gilani, S.O., Kashif, A.S., Jamil, M., Jochumsen, M. and Niazi, I.K., 2021. Evaluation of windowing techniques for intramuscular EMG-based diagnostic, rehabilitative and assistive devices. Journal of Neural Engineering, 18(1), p.016017.

[13]. Fadhli, M., Pedapati, S.R. and Hamdan, H., 2021, May. Development of models for oil and gas pipeline condition prediction using regression analysis. In AIP Conference Proceedings (Vol. 2339, No. 1, p. 020091). AIP Publishing LLC.

[14]. Anghel, V., 2012. Prediction failure for pem fuel cells. International Journal of Advances in Engineering and Technology, 4(2), p.1.

[15]. Sun, H., Burton, H.V. and Huang, H., 2020. Machine learning applications for building structural design and performance assessment: state-of-the-art review. Journal of Building Engineering, p.101816.

[16]. Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W. and Tscheikner-Gratl, F., 2018. Pipe failure modelling for water distribution networks using boosted decision trees. Structure and Infrastructure Engineering, 14(10), pp.1402-1411.

[17]. Mishra, M., Bhatia, A.S. and Maity, D., 2020. Predicting the compressive strength of unreinforced brick masonry using machine learning techniques validated on a case study of a museum through nondestructive testing. Journal of Civil Structural Health Monitoring, pp.1-15.

[18]. Kiani, J., Camp, C. and Pezeshk, S., 2019. On the application of machine learning techniques to derive seismic fragility curves. Computers and Structures, 218, pp.108-122.

[19]. Siam, A., Ezzeldin, M. and El-Dakhakhni, W., 2019, December. Machine learning algorithms for structural performance classifications and predictions: Application to reinforced masonry shear walls. In Structures (Vol. 22, pp. 252-265). Elsevier.

[20]. Mangalathu, S., Jang, H., Hwang, S.H. and Jeon, J.S., 2020. Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls. Engineering Structures, 208, p.110331.

[21]. Hajizadeh, Y., 2019. Machine learning in oil and gas; a SWOT analysis approach. Journal of Petroleum Science and Engineering, 176, pp.661-663.

[22]. Mohamed, A., Hamdi, M.S. and Tahar, S., 2015, August. A machine learning approach for big data in oil and gas pipelines. In 2015 3rd International Conference on Future Internet of Things and Cloud (pp. 585-590). IEEE.

[23]. Hanga, K.M. and Kovalchuk, Y., 2019. Machine learning and multi-agent systems in oil and gas industry applications: A survey. Computer Science Review, 34, p.100191.

[24]. Orrù, P.F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R. and Arena, S., 2020. Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. Sustainability, 12(11), p.4776.

[25]. Layouni, M., Hamdi, M.S. and Tahar, S., 2017. Detection and sizing of metal-loss defects in oil and gas pipelines using pattern-adapted wavelets and machine learning. Applied Soft Computing, 52, pp.247-261.

[26]. Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K., 2019. Modeling tabular data using conditional gan. arXiv preprint arXiv:1907.00503.