COMPARATIVE ANALYSIS AND EVALUATION OF TECHNIQUES FOR GENERATING HIGH-QUALITY SYNTHETIC DATA FOR INDUSTRIAL CONTROL SYSTEMS DATASETS.

by

Gautam Swaminath Ganesh

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
July 2023

Table of Contents

List of Tables

List of Figures

Abstract

In today's society, data plays a vital role in supplying and regulating products that are incoming, outgoing, and currently in play. Machine learning and deep learning techniques based on neural networks have the potential to examine and analyze the data to predict or anticipate certain risks and outputs that could be detrimental to applications. However, these models are not equipped or better facilitate the instances due to a lack of training data, overly biased results (lack of diversity), and are also riddled with privacy concerns since the data used for these predictive algorithms belongs to customers.

In order to address these challenges, the concept of generating synthetic data has shown promising results. The main goal of generating synthetic data is to create artificial data that mimics the statistical characteristics of real-world data and thus can be used for various purposes such as data augmentation, privacy protection, and machine learning model development and evaluation. The primary premise or rule of thumb while generating this synthetic data is to create data that is realistic and representative of the original data while preserving the privacy and confidentiality of sensitive information. Another crucial aspect of generation of synthetic data is the procedure to evaluate it. The prior work has only explored methods to create data that has been tuned and catered to their needs. Thus an unbiased, diverse and standardized method of evaluation is of paramount concern.

This research focuses on the comparative analysis of different methods for generating high-quality synthetic data for Industrial Control Systems (ICS) datasets and proposes an evaluation framework that utilizes visualization and statistical techniques. The main objective is to provide researchers and practitioners with insights into the strengths and limitations of various synthetic data generation approaches and to facilitate the selection of the most suitable method for specific ICS applications. Another objective is to determine synthetic data generation models that could generate high quality diverse datasets that focus on the adverse scenarios and thus making the entire dataset less biased.

The results obtained from the comparative analysis was useful to understand the strengths, limitations and the trade-offs associated with the different synthetic data generation techniques for ICS datasets. The findings obtained from this research helps in identifying the most suitable method to generate high-quality synthetic data for ICS datasets, thereby enabling further research to develop and evaluate algorithms and models in a privacy-preserving environment. This thesis also explores the availability of diverse and realistic synthetic datasets that empower researchers to conduct extensive experiments, validate approaches, and improve the robustness and security of ICS.

Furthermore, this research contributes to the broader field of ICS research by advancing the understanding of synthetic data generation techniques and their applications. Using the comparative analysis and the evaluation can serve to enhance the capabilities of research in developing more efficient and secure ICS.

Acknowledgements

I thank my parents for the opportunities they have provided me. I thank my supervisor, Dr. Sampalli for his guidance and mentorship. I thank my family and my friends who supported me throughout this process.

Chapter 1

Introduction

Industrial Control Systems (ICS) are an essential component of a multitude of industries, namely manufacturing, energy, and transportation. These technologies make it possible to monitor and manage sophisticated industrial machines. ICS are often generated using a combination of sensors, actuators, and controllers. They are often used to gather data, monitor and control operations regarding the business and make informed decisions based on them. The data generated by these ICS are of paramount importance as they are used in a variety of instances such as analysis of historical patterns and trends, development of control algorithms, life-like simulations of various scenarios, and assessment of the impact of potential changes. [13].

1.1 Supervisory Control and Data Acquisition in ICS

SCADA (Supervisory Control and Data Acquisition) networks play a vital role in critical infrastructure to control and monitor the components of industrial control systems (ICSs) remotely in the public and private sectors. Some examples of ICSs include oil and gas refineries, water storage systems, nuclear power plants, manufacturing units, and transportation systems. Such systems use SCADA networks to manage operations remotely such as controlling the flow of gas and oil in oil pipelines, water flow management, monitoring railway tracks, and controlling boilers, solar panels, sensors, and actuators of plant floor machinery [9][1].

SCADA networks are typically divided into three sub-systems: the control center, intermediate sub-SCADA center, and field site component as shown in Figure 1.1 [9]

Industrial Control Systems (ICS) play a critical role in the functioning of various industries, including manufacturing, energy, and transportation. These systems monitor and control complex processes, ensuring the smooth operation of industrial infrastructures. However, the availability of large-scale, high-quality datasets for ICS

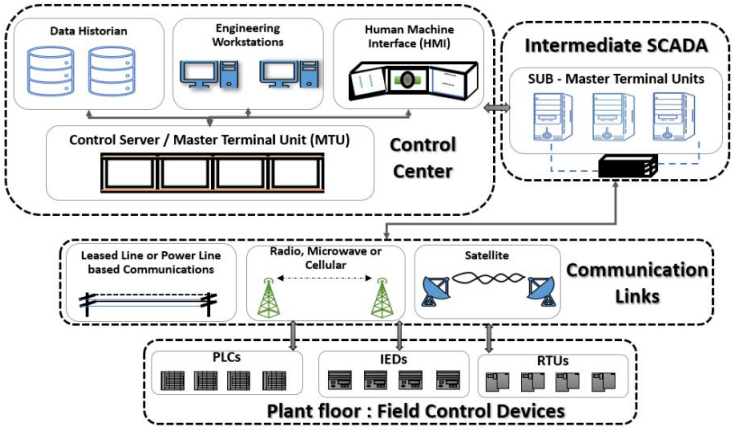

Figure 1.1: Provides an overview of SCADA networks.

research and development is often limited due to privacy concerns, proprietary information, and the potential risks associated with exposing sensitive operational data. To address this challenge, the generation of synthetic data has emerged as a promising approach to creating representative datasets that preserve the statistical characteristics of the original data while ensuring privacy and security [14].

1.2 Data Challenges in Industrial Control Systems

It is difficult to obtain a real-world ICS database for research purposes due to many reasons [10]. Privacy concerns related to the data are one of the major reasons. The sensitive nature of ICS datasets stems from the fact that they include vital data including operating parameters, the method the equipment is set up, and proprietary algorithms. Exposure to such information can lead to detrimental results for the company and could end up aiding rival corporations. Therefore, organizations refuse to disclose their operational data due to information leaks caused by unauthorized access to this data which also includes cyberattacks and weak system security. This results in very less real-world datasets for more study and research to take place. [14].

Additionally, the competitive environment between the corporations in the relevant sectors generally leads to increased difficulty in obtaining ICS datasets. The ICS datasets are some of the tightly guarded intellectual properties of the respective sectors mainly to obtain a competitive advantage. This severely restricts access to various datasets that are of immense value in order to create novel solutions and

expand the field of ICS.

1.3 Synthetic Data Generation for ICS

Few options were available to address this lack of ICS dataset availability [3][36][2]. To prevent data leaking, the researcher must be connected to the organization, or there must be a backup plan in place before the data can be made available to the general public.

Synthetic data creation techniques have become a promising answer to the problems posed by real-world ICS datasets. These synthetic datasets are data that have been manufactured artificially and capture statistical characteristics and trends present in the original dataset, while also omitting any private information. Through this method, researchers can get around the data availability and privacy issues and continue to explore and build solutions using synthetic datasets. [13].

Many alternative approaches have been proposed for the production of synthetic data in many fields. These include Gaussian mixture models (GMMs)[3], variational autoencoders (VAEs)[36], and generative adversarial networks (GANs)[2]. These methods use statistical and machine learning algorithms to create new samples that closely mimic the original data while learning the underlying data distribution from real-world datasets.

1.4 Research Objectives

The primary objective is to the meet the demand for high-quality datasets that provide diverse features and a high level of information preservation. Therefore the creation and assessment of algorithms, models, and privacy-based security measures in the fields of ICS datasets were of prime importance. Using the synthetic data, researchers are able to explore and assess the effectiveness of various algorithms in a controlled environment.

The study's main goal is to compare the various methods by which it is possible to generate synthetic data for ICS datasets. The research also intends to obtain insights into the strengths and limitations along with the applicability of the various synthetic data generation algorithms to various ICS applications and evaluate the performance

and utility of each method. Additionally, a framework for evaluating the usefulness and quality of the generated synthetic data will be created using various statistical analyses, visual analyses, and machine learning algorithms.

By giving researchers access to such high-quality, datasets the results obtained from their research would help to advance the area of ICS research. This will help us to create better algorithms, system architecture as well as better security features for ICS.

1.5 Major Contributions

Due to the selective and private nature of corporations, research into the development of synthetic data generation models intended for ICS datasets and its evaluation are sparse in number and thus have certain research gaps [4][3][35][25]. These are opportunities for development and further improvement in the current study of ICS datasets. The thesis aims to address and suggest solutions for some of these issues.

- Comparative Analysis.
  - Previous research done on synthetic data generation does not encompass a thorough examination of the different ways to generate synthetic data.
- Diversity
  - Although different methods to generate synthetic data are being researched, there is an inherent lack of diversity in this synthetic data.
- Evaluation Metrics
  - Research into synthetic data generation is not complete without evaluating the data and comparing its efficiency when compared with the original data. Therefore it is imperative to have standard evaluation metrics to understand the generated dataset.
- Privacy Preservation
  - A dearth of effective methods to provide adequate confidence in addressing privacy concerns.

- Scalability and Efficiency
  - Research into synthetic data generation for ICS datasets, although large in number still possesses scope for improvement in the fields of increasing the scalability and efficiency.

In this research by identifying and studying these research gaps, we aim to contribute to fields of synthetic data generation by devising a framework for evaluating the synthetic dataset and providing a comparative analysis of different techniques present to generate synthetic datasets. This research also aims to improve the diversity and scalability issues present in the generation of data. Through these contributions, this thesis will aim to advance the methods in generating synthetic data for ICS and also allow for effective evaluation and testing of algorithms in the field of ICS.

1.6 Organization of the Thesis

The following chapters are as follows:

Chapter 2 covers the background and related studies. In this chapter, the past work in the fields of synthetic data generation such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and Gaussian Mixture Models (GMM) are explored, along with that the different evaluation methodologies required to quantify the quality of the synthetic data are also introduced. The challenges faced and the absence or scope of improvement in certain research work are also explored.

Chapter 3 covers the overall methodology of the thesis. It introduces the workflow of the thesis using a flowchart and explains each step. Following up with the different synthetic data generation models and their benefits.

Chapter 4 covers the overall evaluation framework associated with the different synthetic data generation models. Using this framework it is explored on the different methods of evaluating the synthetic data and covers a broad number of aspects. This section also covers a comparative analysis discussing the pros and cons of each model along with the results obtained from the statistical analysis, visualizations and machine learning models.

Chapter 5 summerizes the research with the remarks, limitations and the future work.

Chapter 2

Literature Survey and Related Work

In recent years, the generation of high-quality synthetic data for large Industrial Control Systems (ICS) datasets and its evaluation using visual analytics and statistics have gained significant attention [33][6][32]. With regards to the prior work done by researchers, although the field of synthetic data generation has gained traction and widespread interest in the minds of researchers due to their inherent benefits, research gaps still remain. This section provides an overview of the existing research and methodologies in this domain.

2.1 Literature Survey

In this section, we discuss a deep dive into the multitude of different concepts that will be used to perform the certain tasks such as generation of synthetic data, evaluation of synthetic data using visualizations, statistics and machine learning models. The importance and the previous work related to these concepts are discussed in the chapters ahead.

2.1.1 Challenges faced in Previous Research

Lack of Comparative Analysis.

Although numerous strategies have been put out for creating synthetic data in ICS, there is a dearth of thorough comparisons that rate the efficiency and efficacy of various approaches. A comparison would enable researchers and practitioners to comprehend the benefits, drawbacks, and potential applications of each strategy and choose the best course of action based on the unique specifications and features of the ICS datasets.

Limited Diversity of Generated Data

Numerous research already been conducted and concentrate on creating artificial data that mimics the statistical characteristics of the real dataset but lacks diversity in terms of illustrating various scenarios and variations. For the thorough assessment and testing of algorithms and models in ICS, the provision of different synthetic data that spans a wide range of operational situations, system configurations, and anomalies is crucial.

Lack of Standard Evaluation Metrics

The quality and resemblance of synthetic data to real-world datasets must be evaluated using standardized evaluation measures. It is difficult to compare and benchmark various synthetic data creation approaches since existing research frequently relies on qualitative assessments or domain-specific evaluation metrics carried out primarily by industry professionals. Researchers might analyze and compare the effectiveness of various methods using a similar framework if standardized evaluation metrics were to be developed.

Privacy Preservation

While synthetic data generation techniques aim to address privacy concerns, the effectiveness of these methods in preserving privacy needs further exploration. Research should focus on developing techniques that ensure the privacy and confidentiality of sensitive information in synthetic datasets while maintaining the original data's statistical properties and patterns.

Scalability and Efficiency

Large-scale ICS datasets may need the generation of synthetic data, which can be a time- and resource-intensive process. Techniques that are effective and scalable and can process massive amounts of data while producing synthetic datasets quickly are needed. The development of algorithms and methods that scale to industrial-sized datasets while preserving good data quality should be the main emphasis of research.

2.1.2 Synthetic Data Generation for Industrial Control Systems

Industrial Control Systems (ICS) have been the subject of extensive research into synthetic data creation techniques. With regard to privacy issues, data scarcity, and the requirement for various datasets for testing and assessment reasons, these strategies strive to solve the difficulties associated with restricted access to real-world ICS data. To create synthetic data that accurately replicates the statistical characteristics and patterns of real-world ICS datasets, several methods have been put forth. Here, we examine the literature on creating synthetic data for ICS and emphasize the most important methods and techniques.

Generative Adversarial Networks (GANs)

The production of synthetic data for ICS has drawn a lot of interest in Generative Adversarial Networks (GANs). A generator and a discriminator are the two neural networks that makeup GANs. The discriminator network learns to discriminate between real and fake data, while the generator network learns to create samples of synthetic data. GANs have been used to create synthetic ICS data, and the results are encouraging. In order to create synthetic ICS data for anomaly detection, Zou et al. (2019) [9][6][21][10][29][3][4] used GANs, attaining great fidelity and accurately capturing the underlying statistical characteristics of the original data. A GAN-based method for creating synthetic data in ICS was put forth by Zhang et al. (2020)[30], and it successfully mimicked the distribution and features of real-world data. By learning from the underlying data distribution, GANs offer a potent framework for producing high-quality synthetic data.

Autoencoders

Autoencoders are neural network architectures used for unsupervised learning tasks, including synthetic data generation. Autoencoders consist of an encoder network that maps the input data to a lower-dimensional latent space and a decoder network that reconstructs the input data from the latent space. Autoencoders have been employed in the generation of synthetic ICS data for various purposes, including anomaly detection and system modeling. Li et al. (2018) [14][7][16][5] proposed a

synthetic data generation method using autoencoders for anomaly detection in ICS. The autoencoder effectively captured the underlying patterns and anomalies in the original data, providing high-quality synthetic samples. Autoencoders offer flexibility in capturing the complex relationships present in ICS data and generating realistic synthetic samples.

Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are generative models that combine the concept of autoencoders with probabilistic modeling [33][6][15][32][24][37]. VAEs learn a latent space representation of the input data, allowing for the generation of new samples by sampling from the learned distribution. VAEs have been utilized in the generation of synthetic ICS data, enabling the capture of complex data distributions and generating diverse samples. Wang et al [18]. (2020)[16][5] proposed a VAE-based approach for synthetic data generation in ICS, which effectively captured the statistical properties and patterns of the original data. VAEs provide a probabilistic framework for generating synthetic data with controllable characteristics, making them suitable for diverse applications in ICS.

Gaussian Mixture Models (GMM)

Synthetic data generation using Gaussian Mixture Models (GMMs) has gained attention in the domain of Industrial Control Systems (ICS) datasets [26][27]. Researchers have explored the application of GMM models to generate realistic and diverse synthetic data that mimics the characteristics of ICS datasets. For example, Deka et al[34] . (2019) proposed a GMM-based approach for generating synthetic ICS network traffic data, which captured the statistical properties and temporal dependencies of real-world ICS network traffic. The study by Yang et al [18]. (2020) utilized GMM models to generate synthetic ICS sensor data, considering the multivariate nature and complex dependencies present in the original dataset. These studies demonstrated the potential of GMM models in generating synthetic data that closely resembles the statistical properties of real ICS datasets.

2.1.3 Evaluation of Synthetic Data

Synthetic data is only considered useful when it has been carefully evaluated to better understand its quality and utility for ICS applications. Visual analytics and statistical techniques play a major role in assessing and evaluating newly generated synthetic datasets. [23]

Visual analytics techniques provide an exploratory analysis of the synthetic dataset. Chen et al. [13] proposed a visual analytics framework that enables users to visually compare the synthetic and real data, identify discrepancies, and evaluate the effectiveness of the generation model.

In order to determine and conform if the statistical properties and distribution of data is performed to a similar degree as the original dataset, it is imperative to utilize the various methodologies provided by statistical analysis. Johnson et al.[20][14] presented a methodology to evaluate the synthetic datasets using measures such as mean, variance and correlation. By evaluating such metrics based on the original dataset, the study was able to understand and quantify the properties of the synthetic datasets.

2.1.4 Evaluation Using Statistics

One approach for evaluating synthetic data is through the use of statistical metrics. Zhang et al [30]. (2016) proposed the use of the Kolmogorov-Smirnov distance and the Wasserstein distance to measure the similarity between the statistical distributions of real and synthetic data. They demonstrated the effectiveness of these metrics in evaluating the quality of synthetic data generated for healthcare datasets. Similarly, Salem et al [10] introduced the concept of Precision-Recall Curves to evaluate the utility and privacy preservation of synthetic data generated for social network datasets. These statistical metrics provide quantitative measures of the similarity and accuracy of the synthetic data compared to the original data.

They demonstrated the effectiveness of these metrics in evaluating the accuracy of synthetic data generated for power systems datasets. Similarly, Hodo et al [24][28]. (2017) introduced a set of metrics, including correlation, entropy, and histogram comparisons, to evaluate the similarity and preservation of statistical characteristics in synthetic ICS datasets.

2.1.5 Evaluation Using Visual Analytics

With the results presented by the statistical analysis, in order to supplement the regions that they lacked an alternative approach is taken into consideration, this being visual analytics. Wong et al [37][28] introduced the concept of Visual Data Comparison (VDC) to visually compare the synthetic data with the original data. Scatter plots, box plots, and heatmaps are some of the most common methods used in order to analyze and view the distribution. This allowed for a more comprehensive view of the synthetic data and provided information about the quality of the dataset.

Furthermore, visualization techniques have been utilized for the evaluation of synthetic data in the ICS domain. Monteiro et al[37][21] introduced a visual analytics framework to assess the quality and usefulness of synthetic ICS datasets. They employed scatter plots, heatmaps, and time series visualizations to compare the distributions, correlations, and temporal patterns of the synthetic and real data. This visual analysis allowed for the identification of various discrepancies and anomalies in the synthetic data, providing insights into its quality and applicability.

2.1.6 Evaluation Using Machine Learning Models

The evaluation of synthetic data can also involve the use of machine learning models. Liu et al [30] proposed an evaluation framework based on the classification performance of a machine-learning model trained on synthetic data. They compared the accuracy, precision, and recall of the model when trained on synthetic data versus real data, demonstrating the effectiveness of their evaluation approach.

Another important aspect of evaluating synthetic data for ICS datasets is assessing the information preservation capabilities. Giraldo et al[12] proposed an evaluation framework based on the comparison of system behavior using dynamic simulators. They compared the behavior of the original and synthetic datasets by running simulations and analyzing the system response. This approach allowed for the assessment of the accuracy and realism of the synthetic data in representing the dynamics of the ICS.

2.1.7 Quality Metrics for Synthetic Data

Several metrics have been proposed to measure the quality of synthetic data generated for large ICS datasets. One widely used metric is the Wasserstein distance, which quantifies the dissimilarity between the distributions of real and synthetic data. Xu et al[10] utilized the Wasserstein distance to evaluate the quality of synthetic ICS data and demonstrated its effectiveness in capturing the distributional differences. Another important metric is information loss, which assesses the extent to which the synthetic data preserves the original information from the real dataset. Nguyen et al[30] proposed an information loss metric that measures the dissimilarity between the original and synthetic data in terms of their mutual information. This metric enables researchers to quantify the loss of information during the generation process.

2.1.8 Privacy-Preserving Synthetic Data Generation

Privacy concerns are of utmost importance when dealing with sensitive ICS datasets. To address this, privacy-preserving synthetic data generation techniques have been proposed. For example, Li et al[38] introduced a differential privacy mechanism in the synthetic data generation process to ensure individual privacy protection. By adding controlled noise to the generated data, the privacy of individuals in the original dataset is preserved while maintaining the statistical properties.

2.1.9 Real-World Applications

The generation of high-quality synthetic data for large ICS datasets has found applications in various domains. For instance, Zhang et al[27] utilized synthetic data to enhance anomaly detection in ICS networks. By augmenting the real dataset with synthetically generated samples, they achieved improved detection accuracy and robustness against novel attacks. Another application is the training of machine learning models on synthetic data to overcome the limitations of limited real-world datasets. Dong et al[19] proposed a deep learning approach that leverages synthetic data for training intrusion detection systems in ICS. Their results demonstrated the effectiveness of synthetic data in improving the performance and generalization of the detection models. The research on generating high-quality synthetic data for large

Industrial Control Systems datasets and evaluating them using visual analytics and statistics is a rapidly evolving field. The use of generative models, such as GANs and VAEs, combined with visual analytics and statistical techniques, holds great potential for generating realistic and representative synthetic data. Privacy-preserving mechanisms and quality metrics further contribute to the advancement of synthetic data generation. Real-world applications demonstrate the practical value of synthetic data in enhancing various aspects of ICS security and analysis.

2.2 Analysis on Related Work

In this section, the core research studies and their advantages and disadvantages are studied and tabulated. These studies are chosen based on relevance, types of synthetic data generation model, previous studies on different evaluation methodologies and comparative analysis.

S.No	Title	Type of Model	Positives	Detriments
1.	A Statistical Framework for Evaluating Synthetic Datasets for Industrial Control Systems [36]	GAN model, CTGAN model	Synthetic data generation model performs amicably well. Statistical tests were performed to validate statistical accuracy.	Distribution tests were not performed. Does not possess in-depth evaluation into all factors of synthetic data generation. Do not account for data diversity.
2.	Synthetic ICS Datasets Generation Framework for Training Intrusion Detection Systems [35]	GMM model	Synthetic data generation model performs amicably well. In-depth statistical analysis was performed.	Distribution tests were not performed. Visualization is not utilized for evaluation. Does not account for data diversity.
3.	Visualization Techniques for Evaluating Synthetic Industrial Control System Datasets [29]	VAE model	An indepth analysis using visualization techniques were utilized	Tests were not performed for a variety of datasets. Does not account for data diversity.
4.	Statistical Evaluation of Synthetic Power System Data. [25]	GAN model	Synthetic data generation model performs amicably well. Statistical tests were performed to validate statistical accuracy. An in-depth statistical analysis was performed.	Distribution tests were not performed. Does not possess in-depth evaluation into all factors of synthetic data generation. Do not account for data diversity. Visualization is not utilized for evaluation.
5.	A Visual Analytics Approach for Synthetic ICS Data Evaluation [3]	GAN model	Various visualization techniques were utilized. In-depth test into the distribution of data performed.	Diversity of Synthetic Data not explored.
6.	Metrics for Evaluating the Quality of Synthetic Intrusion Detection System Datasets [4]	VAE model	Diversity of synthetic data is considered. Statistical accuracy and privacy aspects are explored	Does not include visual analytics to understand the distribution of synthetic data.

Table 2.1: Comparison of Evaluation Techniques

Chapter 3

Methodology of Research

In this chapter, we present the methodology employed to carry out our research. This section will be divided into the different components that substantiate this thesis starting with the dataset that is used, different models that were used, the preprocessing and postprocessing methods used, the smoothening algorithms that were tested, and the different evaluation algorithms that were used to perform the comparative analysis on the ICS dataset.

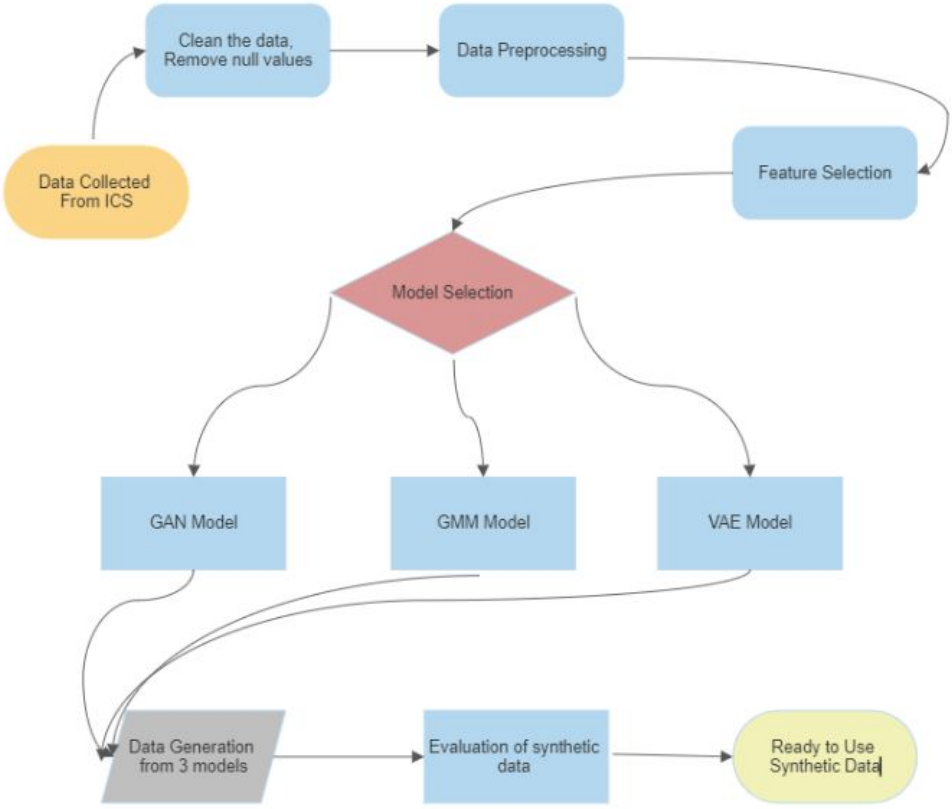

Figure 3.1: Flow chart representing the overall methodology

The overall workflow is depicted in the above flowchart. The ICS dataset is subjected to a thorough preprocessing algorithm. This preprocessing algorithm consists

of searching for null values and removing error values, and a rigorous feature selection process. The preprocessing algorithm outputs the top 10 features in the dataset based on correlation and feature importance. Once the data is then cleaned and preprocessed it is then fed to one of the three synthetic data generation algorithms namely the GAN model, the GMM model, and the VAE model. The synthetic data that is obtained at the end of these algorithms are not ready to use as they must undergo a postprocessing algorithm that consists of methods to convert the data into a more usable format.

Having access to a reliable gas pipeline ICS dataset enables researchers to develop and validate predictive models and algorithms that can detect and prevent catastrophic events, such as pipeline ruptures or explosions. This data plays a central role in the generation of synthetic data using the various machine learning algorithms and eventually, the evaluation of the said synthetic data thus providing valuable information.

3.1 Dataset Used: Gas Pipeline ICS dataset

The data that is used is the Pipeline dataset, this dataset encompasses the different features that determine if a pipeline would explode or not. The significance of this dataset is that number of times statistically a pipeline would explode in real life is significantly less due to the improvements in current-day technology. Thus the number of instances where the dataset concludes that the pipe would explode is far smaller than its counterpart. Thus Synthetic data here plays a major role in accomplishing this goal. It reduces the innate bias present in the data and also augments the dataset.

This ICS dataset comprises multiple sensors and actuators that help determine the current state of the pipes. These sensor values can help to analyze and predict the state of the pipe and if it would stay in a benign state or not. The availability of a comprehensive and accurate gas pipeline Industrial Control Systems (ICS) dataset is of paramount importance when it comes to ensuring the safety and reliability of gas transportation infrastructure. Such a dataset plays a crucial role in assessing the risks associated with the operation of gas pipelines and in determining the likelihood of potential explosions or other hazardous incidents. By analyzing and modeling the data, researchers can gain insights into the various factors that contribute to pipeline

integrity and safety, including pressure levels, temperature variations, flow rates, and system anomalies. The Table bellow provides a description into the different features in the dataset.

Feature	Description
response_address	refers to the destination address to which a response is sent or expected
resp_length	refers to the length of the response message in a communication transaction. It represents the number of bytes or bits contained in the response payload.
response_memory	refers to the memory consumption or utilization associated with the response message in a communication transaction.
resp_write_fun	The resp_write_fun field captures information about the function or method responsible for writing the response message.
response_memory_count	refers to the count or number of memory operations associated with a response message.
resp_read_fun	refers to the type or function associated with a read operation performed as part of generating a response message.
setpoint	refers to a predetermined or desired value that is used as a reference for controlling a specific process or system parameter. It represents the target value or desired state that the system aims to achieve and maintain.
control_mode	It indicates the specific control algorithm or method employed to regulate and maintain the desired parameters within the pipeline.
pump	provides information about the characteristics and behavior of the pump within the pipeline system.
command_address	refers to the address associated with a command sent to a certain component within the pipeline system
command_memory	refers to the memory location or data storage area where a command is stored
command_memory_count	refers to the number of commands or instructions stored in the memory of the control system
comm_read_function	refers to the function or method used for reading or retrieving data from the communication channel or protocol within the control system.
comm_write_fun	refers to the function or method used for writing or sending data to the communication channel or protocol within the control system.
sub_function	contributes to the overall control and operation of the ICS pipeline.
command_length	It represents the number of characters, bytes, or bits that make up the command.
crc_rate	refers to the CRC (Cyclic Redundancy Check) error rate.
time	refers to the time taken in the opperation
result	refers to the outcome of the particular instance

Table 3.1: Represents the different parameters in the pipeline dataset

3.2 Synthetic data generation

3.2.1 Machine Learning Model 1: Generative Adversarial Networks (GANs)

GAN models or Generative Adversarial Network (GAN) model for synthetic data generation. The GAN consists of two neural networks - a generator and a discriminator. The generator creates new synthetic samples, while the discriminator evaluates the quality of the generated samples by comparing them to the real data [22]. There are many classifications within GAN Models, in this case, CTGAN models will be taken into consideration over the normal GAN models for a few reasons.

Both GANs (Generative Adversarial Networks) and CTGAN (Conditional Tabular GANs) are generative models that can be used to create synthetic data that mimics the statistical properties of real data. However, there are some key differences between the two:

- Input data format: GANs can work with any type of data, including images, text, and sound. CTGAN, on the other hand, is designed specifically for tabular data, which is data that is organized into rows and columns.
- Conditioning: CTGAN can generate synthetic data that is conditioned on a set of input features, whereas GANs typically generate data without explicit conditioning. This means that CTGAN can be used to generate synthetic data that is more similar to a specific subset of the real data.
- Training data: CTGAN requires labeled training data, whereas GANs can be trained using unlabeled data. This means that CTGAN is more suitable for supervised learning tasks, where there is a clear distinction between input and output variables.
- Output: GANs generate synthetic data that is optimized to be as realistic as possible, whereas CTGAN generates synthetic data that is optimized to be as similar as possible to the real data, while also satisfying any conditioning constraints.

Overall, CTGAN is a specialized type of GAN that is designed specifically for tabular data, with the ability to generate synthetic data that is conditioned on input features.

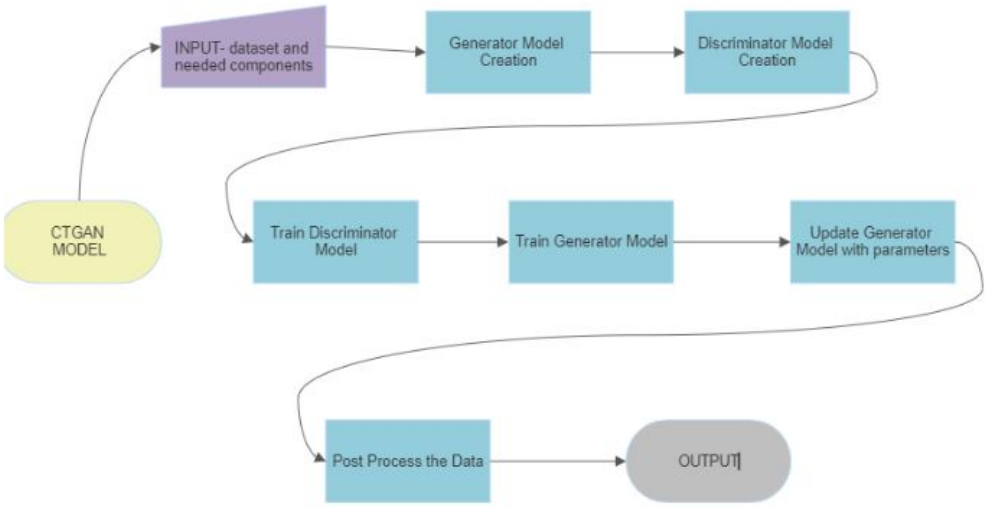

Figure 3.2: Represents the generation of synthetic data using the CTGAN model

In this model, we initialize certain parameters before preceding through with the CTGAN model. The dataset, the number of epochs, batch size, the update rate of the discriminator and the generator, and finally the learning rate. The Overall workflow of the CTGAN model follows as per Fig.3 .

In this model, the data initially undergoes preprocessing in the form of converting categorical columns into one-hot encoded columns and simultaneously another process is also undertaken to normalize the continuous columns. This is done in order to better understand the data and process it using the generators and the discriminator models. Subsequently, the next task at hand is to initialize the generator models and the discriminator models. The generator model is a neural network with multiple fully connected layers that takes a noise vector as input and outputs a synthetic sample in contrast the discriminator model is a neural network with multiple fully connected layers that takes a real or synthetic sample as input and outputs a binary classification (real or fake).

In order to train the discriminator, initially feed a batch of real and synthetic data samples to the discriminator. Once the data has been fed compute the binary cross entropy loss between the true labels (1 for real, 0 for fake) and predicted probabilities. Once that is complete update the discriminator model using backpropagation and use

the Adam optimizer for accurate and stable results. Repeat the above process for a fixed number of epochs.

Likewise, in order to train the generator. A batch of noise vectors is initially generated. These noise vectors are then subsequently fed to the generator to obtain a batch of synthetic data samples. With the synthetic data samples at hand, feed these samples to the discriminator model in order to obtain the predicted probabilities. Compute the binary cross-entropy loss between the true labels (1 for real, 0 for fake) and predicted probabilities (but swap the labels, i.e., use 1 for fake). Once this process is complete update the generator model using backpropagation and the Adam optimizer and repeat the above process for a fixed number of epochs.

Once the generator and the discriminator models have been trained and a batch of resulting synthetic data is obtained, a postprocessing algorithm is applied. This process includes the de-normalization of all the continuous columns and the conversion of the one-hot encoded categorical columns back to categorical columns. With this complete, the resulting output is then evaluated in order to check its performance and likeliness to the original dataset.

With the complete iteration of the CTGAN model, the output obtained is synthetic data that closely resembles the original dataset. This synthetic data are trained in such a manner that it closely resembles the original dataset in terms of statistical similarity and distribution as well [19]. The CTGAN (Conditional Tabular GAN) model offers several benefits in the generation of synthetic data for Industrial Control Systems (ICS) datasets compared to other synthetic data generation models [10]. Some of the key benefits include:

- CTGAN models are used specifically for tabular data. This property makes it suitable for ICS datasets which are in the form of tabular datasets. The CTGAN models unlike the other models are tailored in such a way that it captures the characteristics and dependencies present in the ICS datasets.

- CTGAN models are great at generating synthetic data that possess identical if not close to the original datasets' conditional relationships and dependencies. This is a major factor for ICS datasets as they are mainly comprised of interdependencies within the multitude of variables.

- CTGAN models provide a flexible method by incorporating domain knowledge of conditional variables. This allows for further study that could be done in analyzing the attributes or conditions pertaining to the ICS dataset, therefore, resulting in a context-aware synthetic data generation process.

- CTGAN models employ a large number of deep learning techniques and optimization algorithms to generate synthetic data that is capable of handling large-scale synthetic data This makes it suitable for large ICS datasets in the real world.

- CTGAN models are able to generate highly realistic synthetic data that can mimic the statistical properties and distributions of the original dataset. CTGAN models, upon learning from the original dataset can replicate the underlying patterns and characteristics present in the synthetic data.

- Evaluation and Validation: CTGAN provides built-in evaluation metrics to assess the quality of the generated synthetic data. These metrics can be used to compare the statistical properties and distributions between the original and synthetic datasets, enabling researchers to validate the effectiveness of the CTGAN model for generating realistic and reliable synthetic data.

3.2.2 Gaussian Mixture Models (GMM)

The subsequent model in question is the Gaussian Mixture Model. The overall workflow of the model can be seen with the assistance of Fig.4 .

Figure 3.3: Represents the generation of synthetic data using the GMM model

In this model, the data undergoes a preprocessing step which includes the collection of the ICS dataset and performing a cleaning operation, which consists of removing outliers, handling missing values, and normalizing the data. A feature selection process is also done in order to identify all relevant features from the preprocessed

dataset that capture the essential characteristics of the ICS system. The features are then selected based on their significance and relevance in accordance with the specific analysis or modeling objective. This process is predominantly done using correlation and covariance. The selected features are then compiled and fed to the GMM model.

The Gaussian Mixture Model algorithm is then applied to the preprocessed dataset in order to understand and analyze the underlying probability distribution of the data. Once this is complete, the appropriate number of Gaussian components in the GMM is to be determined. This is accomplished using techniques such as the Bayesian Information Criterion (BIC). Once this is complete, estimate the parameters of the GMM, including the means, covariance matrices, and mixture weights, using the expectation-maximization (EM) algorithm.

In order to generate the synthetic dataset, a sampling operation is performed on the GMM model. Using the learned parameters from the above tests synthetic data samples are obtained from the trained GMM model. The parameters that were obtained using the BIC and EM algorithms are used to make the synthetic data resemble the original ICS dataset. The GMM model then generates samples by randomly selecting a component according to the mixture weights and sampling from the corresponding Gaussian distribution.

The synthetic data that is obtained is then adjusted to meet specific requirements or constraints and data transformations or adjustments to align the data with the characteristics of the real ICS datasets are performed. Once the synthetic data is obtained it is imperative to evaluate the quality of the generated synthetic data using various metrics, such as the similarity between the real and synthetic datasets, statistical measures, and domain-specific evaluation criteria.

Once the evaluation is underway a comparison of the statistical properties of the synthetic data with the original ICS dataset to ensure the synthetic data captures the essential characteristics and maintains the integrity of the system. If the quality of the synthetic data does not meet the desired criteria, iterate and refine the GMM model and generation process. Adjust the GMM parameters, such as the number of components, initialization strategy, or covariance structure, and repeat the training and sampling steps. Using the above steps as a roadmap, continuously evaluate the generated synthetic data and refine the process until satisfactory results are achieved.

It is important to note that the benefits of the GMM model may vary depending on the specific characteristics and requirements of the ICS dataset [32]. Other synthetic data generation models, such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs), may also offer unique advantages in certain scenarios. Therefore, the selection of the most suitable synthetic data generation model should consider the specific context, objectives, and characteristics of the ICS dataset under investigation [18][5].

The Gaussian Mixture Model (GMM) has several benefits in the generation of synthetic data for Industrial Control Systems (ICS) datasets compared to other synthetic data generation models [18]. Some of the key benefits include:

- Capturing Complex Data Distributions: GMM is capable of capturing complex data distributions by modeling the data as a mixture of multiple Gaussian components. This allows GMM to represent the underlying structure of the ICS dataset more accurately, especially when the dataset exhibits multi-modal or non-linear distribution patterns.

- Flexibility in Modeling: GMM provides flexibility in modeling the data distribution by allowing the specification of the number of Gaussian components. This flexibility allows the GMM model to adapt to different ICS datasets with varying complexities and capture the inherent variations in the data more effectively.

- Generation of Realistic Samples: GMM can generate synthetic samples that closely resemble the original data distribution. By learning the parameters of the Gaussian components from the real ICS dataset, GMM can produce synthetic data points that preserve the statistical properties, correlation structure, and patterns observed in the original dataset.

- Fine-grained Control over Generated Data: GMM allows fine-grained control over the generated data through the manipulation of the mixture weights and parameters of individual Gaussian components. This control enables researchers to influence the characteristics of the synthetic data, such as adjusting the noise level and skewness level or generating samples with specific characteristics.

- Scalability: GMM can handle large-scale ICS datasets efficiently due to its ability to estimate the parameters using the Expectation-Maximization (EM) algorithm, which is computationally efficient for high-dimensional data. This scalability makes GMM suitable for generating synthetic data for large-scale ICS datasets encountered in real-world industrial systems.

- Interpretable Results: GMM provides interpretable results as it assigns a probability to each data point indicating its association with a particular Gaussian component. This information can be useful in understanding the generated synthetic data and assessing its similarity to the real ICS dataset.

3.2.3 Variational Autoencoders (VAEs)

The final model in consideration in this research thesis is the Variational Autoencoders (VAEs). The overall workflow is shown in Fig.5.

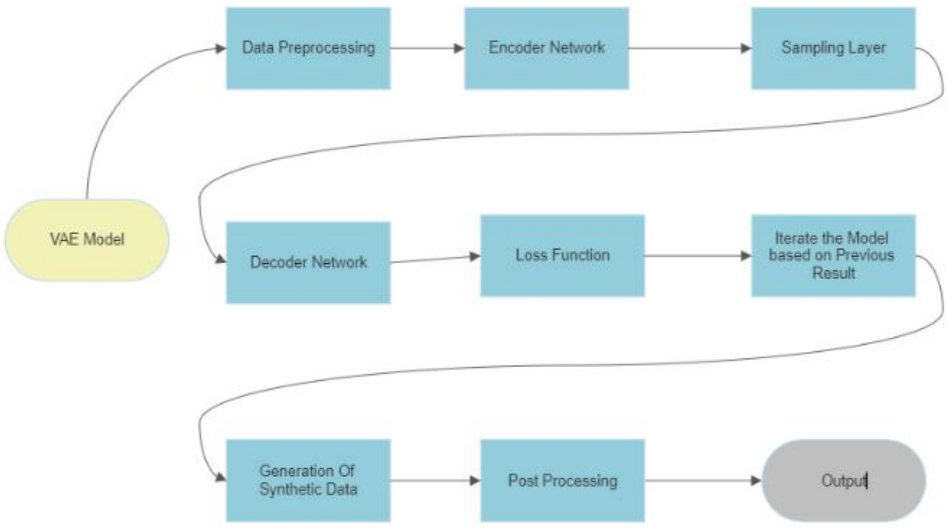

Figure 3.4: Represents the generation of synthetic data using the VAE model

In this model the input original ICS dataset is initially preprocessed, ie. the data is normalized between 0 and 1. This data is then split into training and testing datasets. Once this is complete, the Encoder network is to be initialized.[29]

The encoder network is designed to have an input layer to receive the data, several

hidden layers present to understand the representation of the data, and finally, an output layer that represents the latent space z with a mean vector and a standard deviation vector. The encoded data is then fed to the Sampling Layer, where the sample is generated using the reparameterization technique using the mean and the standard deviation vectors.

Subsequently, the decoder network is also initialized, by having an input layer that would receive the sampled data of a particular latent space, several hidden layers to reconstruct the data, and an output layer to generate the synthetic data. There are two different loss functions used in these networks. The First one is the reconstruction loss function which denotes the difference between the input and the output data. And the KL Divergence loss denotes the difference between the learned distribution and the prior distribution.

Once the decoder is then initialized, the training process begins. This involves backpropagating the loss through the model, this is performed in order to update the weights. This process is repeated until the loss function is minimized. Generate the sample of a particular latent space from a standard normal distribution and use the decoder network to generate synthetic data from the sampled latent space. The synthetic data that is obtained from this process is then evaluated using various metrics and the process is then repeated in order to achieve the desired quality.

Overall, the VAE model is an unsupervised learning model that can learn the underlying structure of the input data and generate synthetic data that is similar to the input data [13]. The model can be trained on a variety of input data types, including images, audio, and text. The VAE (Variational Autoencoder) model offers several benefits in the generation of synthetic data for Industrial Control Systems (ICS) [1] datasets compared to other synthetic data generation models. Some of the key benefits include:

- Latent Space Representation: VAEs learn a latent space representation of the data, which allows for the continuous and structured generation of synthetic samples. This means that the VAE can generate data points in a way that smoothly interpolates between different samples, providing more diversity and flexibility in the generated synthetic data.

- Probabilistic Generation: VAEs model the data distribution using probabilistic techniques, allowing for the generation of synthetic data that captures the inherent uncertainty and variability present in the original ICS dataset. This probabilistic framework enables more realistic and robust synthetic data generation.

- Encoder-Decoder Architecture: VAEs consist of an encoder network that maps the original data into a lower-dimensional latent space, and a decoder network that reconstructs the data from the latent space. This architecture enables the VAE to capture the complex dependencies and patterns present in the ICS dataset, resulting in more accurate and realistic synthetic data generation.

- Continuous and Disentangled Latent Representations: VAEs encourage disentangled representations in the latent space, meaning that different dimensions of the latent space capture independent and meaningful factors of variation in the data. This property allows for better control over the generation process, as specific features or attributes can be manipulated in the latent space to generate synthetic samples with desired characteristics.

- Ability to Learn Complex Distributions: VAEs are capable of learning complex data distributions, including multimodal distributions, which are often encountered in ICS datasets. This capability ensures that the generated synthetic data captures the diversity and variability of the original dataset, leading to more accurate representations of real-world scenarios.

- Reconstruction-Based Evaluation: VAEs can be evaluated based on the quality of the reconstructed data. By comparing the original and reconstructed samples, researchers can assess the fidelity and accuracy of the VAE model in capturing the characteristics of the ICS dataset. This evaluation provides insights into the reliability and effectiveness of the VAE for synthetic data generation.

- Interpretable Latent Space: VAEs can provide interpretable representations in the latent space, where each dimension corresponds to a meaningful attribute or feature of the data. This interpretability facilitates better understanding and

analysis of the generated synthetic data, as researchers can examine and manipulate specific dimensions to explore different aspects of the data distribution.

With the generation of synthetic data using the three machine learning algorithms namely CTGAN, GMM, and VAE models complete and usable synthetic data has been generated, and a thorough evaluation process will now commence.

Chapter 4

Evaluation of Synthetic Data

One of the main objectives of this thesis is to establish a structured evaluation pattern by which one could formulate and understand the statistical patterns and trends along with an extensive evaluation of newly generated synthetic data and how well it performs when compared with the original ICS datasets. Along with this, it is also imperative to provide a comparative analysis of the different means of generating synthetic data and displaying their characteristics using various means, namely using Statistical analytics, Visual analytics, and Machine learning algorithms to provide an in-depth view of the synthetic data.

Evaluating the quality of synthetic data is crucial to ensure its reliability, usefulness, and applicability in various domains and applications. When assessing the quality of synthetic data, it is essential to consider multiple categories or criteria. These categories provide a comprehensive framework for evaluating different aspects of synthetic data and enable researchers and practitioners to make informed decisions. In this section, we will discuss the key categories that one must check to evaluate the quality of synthetic data.

- Fidelity: Fidelity refers to the degree to which the synthetic data accurately represents the original dataset [26]. It involves assessing how well the statistical properties, distributions, and relationships of the original data are preserved in the synthetic data. Measures such as mean, variance, covariance, and higher-order statistics can be used to compare the statistical properties between the original and synthetic datasets. Additionally, visual inspection, hypothesis testing, and domain experts' judgment can provide insights into the fidelity of the synthetic data.

- Privacy and Information Preservation: Privacy preservation is a critical consideration when dealing with synthetic data, especially when the original data

contains sensitive or personal information. Evaluating privacy preservation involves assessing how well the synthetic data obscures or anonymizes personal attributes or sensitive data [17]. Metrics such as k-anonymity, l-diversity, and t-closeness can be used to measure the level of privacy protection provided by synthetic data. Additionally, techniques such as differential privacy can be applied to ensure privacy guarantees. Using these methods we can also check if sensitive and vital information is preserved in the newly generated synthetic data.

- Diversity and Generalization: Diversity measures the variety and coverage of the synthetic data across different attributes and feature combinations. It assesses whether the synthetic data captures the inherent variability present in the original dataset. Metrics such as entropy, distinctiveness, or clustering techniques can be used to measure the diversity of the synthetic data [11]. A diverse synthetic dataset ensures that it represents the full range of patterns and characteristics present in the original data. Generalization assesses how well the synthetic data generalize to unseen or out-of-sample data. It measures the robustness of the synthetic data generation method in capturing the underlying patterns and characteristics of the original dataset. It is important to ensure that the synthetic data accurately represents the real-world distribution of the data to ensure reliable and robust generalization.

- Interpretability and Utility: Interpretability refers to the understandability and transparency of the synthetic data generation process. It involves assessing whether the synthetic data can be easily interpreted and analyzed by domain experts [8]. Techniques such as feature importance ranking, feature engineering, or model interpretability methods can be applied to enhance the interpretability of synthetic data. Utility refers to the usefulness and applicability of synthetic data in specific tasks or applications. It involves assessing how well the synthetic data performs in downstream tasks, such as predictive modeling, classification, or clustering. The performance of models trained on synthetic data can be compared to models trained on the original data to measure the utility of the synthetic data.

In this chapter, we will be exploring these subtopics in detail in order to determine and understand their usage and explore the results of the synthetic data generated by the various machine learning model namely the CTGAN, GMM, and VAE models.

4.1 Fidelity

Fidelity is a critical aspect when evaluating the quality of synthetic data. It refers to the extent to which the synthetic data accurately represents the statistical properties, distributions, and relationships present in the original dataset. In other words, fidelity measures how well the synthetic data captures the essence of the original data [26].

To assess fidelity, various statistical measures and techniques can be employed. Here are some key considerations when evaluating fidelity in synthetic data:

- Statistical Properties: Fidelity evaluation involves comparing the statistical properties of the original and synthetic datasets. This includes measures such as mean, variance, covariance, skewness, and kurtosis. Statistical tests, such as t-tests, can be applied to determine if there are significant differences between the statistical properties of the original and synthetic data [26]. A high degree of similarity in statistical properties indicates a higher fidelity of the synthetic data.

- Distribution Matching: Synthetic data should ideally replicate the distributions observed in the original data. This includes capturing the shape, spread, and tail behavior of the distributions. Visual inspection, kernel density estimation, or quantile-quantile plots can be used to compare the distributions of original and synthetic data [26]. Close alignment between the distributions indicates a higher fidelity of the synthetic data.

It is important to note that achieving perfect fidelity between the synthetic and original datasets may not always be feasible or necessary. The goal is to strike a balance between fidelity and privacy protection. Synthetic data may intentionally introduce some level of noise or perturbation to preserve privacy, which may result in slight deviations from the original data. The level of fidelity required will depend on the specific application and the trade-off between privacy and utility.

Overall, fidelity assessment provides insights into the accuracy and reliability of the synthetic data. By ensuring a high level of fidelity, researchers and practitioners can have confidence in using synthetic data for various analyses, modeling, and decision-making processes.

4.1.1 Descriptive Statistics

One of the primary and important techniques that could be used to determine the statistical properties is descriptive statistics. Descriptive statistics take into account the mean, mode, range, variance, and standard deviation [33][28]. Using this it is possible to determine whether the statistical properties possessed by the synthetic data are close to the original dataset.

Para	res_add	resp_lgt	res_mem	res_wrt_fun	res_mem_cnt	res_rd_fun	setpoint	ctrl_md	pump	result
Count	97019	97019	97019	97019	97019	97019	97019	97019	97019	97019
Mean	3.719	26.29	216.65	9.298	16.737	2.44	24.16	0.899	0.056	0.36
Std	1.021	26.56	59.5	2.55	4.59	0.893	14.322	0.991	0.23	0.482

Table 4.1: represents the descriptive statistics output for the original dataset

Para	res_add	resp_lgt	res_mem	res_wrt_fun	res_mem_cnt	res_rd_fun	setpoint	ctrl_md	pump	result
Count	97019	97019	97019	97019	97019	97019	97019	97019	97019	97019
Mean	1.14	48.77	66.48	2.88	5.15	1.98	39.77	0.97	0.953	0.053
Std	0.512	12.12	27.07	1.19	2.11	0.193	8.101	0.17	0.211	0.224

Table 4.2: represents the descriptive statistics output for the CTGAN dataset

Para	res_add	resp_lgt	res_mem	res_wrt_fun	res_mem_cnt	res_rd_fun	setpoint	ctrl_md	pump	result
Count	97019	97019	97019	97019	97019	97019	97019	97019	97019	97019
Mean	3.714	26.29	216.4	9.286	16.71	2.34	24.26	0.899	0.033	0.37
Std	1.029	26.77	59.99	2.574	4.63	0.903	14.509	1.026	0.238	0.483

Table 4.3: represents the descriptive statistics output for the GMM dataset

Para	res_add	resp_lgt	res_mem	res_wrt_fun	res_mem_cnt	res_rd_fun	setpoint	ctrl_md	pump	result
Count	97019	97019	97019	97019	97019	97019	97019	97019	97019	97019
Mean	3.196	40.73	185.2	7.621	14.24	2.649	34.318	0.589	0.0047	0.0053
Std	0.652	14.38	32.32	1.382	2.565	0.490	9.645	0.539	0.211	0.2249

Table 4.4: represents the descriptive statistics output for the VAE dataset

4.1.2 Mahalanobis Distance

From [33] Mahalanobis distance is a statistical measure that quantifies the distance between a point and a distribution of points in a multivariate space. It considers the correlations between variables, which makes it particularly useful when dealing with datasets with multiple variables that are not necessarily independent The Mahalanobis distance is a normalized measure that considers the variances and covariances of the variables in the dataset. It is essential in evaluating the similarity between an observation and the distribution it belongs to. This test allows one to check the authenticity of the synthetic dataset. Using this test it can be found whether the dataset is from the same distribution or from a different one and also how different the dataset is when compared with the original dataset.

The Mahalanobis distance calculation can be divided into three main steps. Initially, the data is preprocessed by centering and scaling it. This can be done by subtracting the mean value from each variable and further dividing it by the standard deviation. This method ensures that all variables have been scaled and there are no overrepresentations or dominance by variables with larger value ranges. THe next step is the calculation of the covariance matrix in order to capture the relationships and correlations between variables. If there are n variables, the covariance matrix will be an n x n matrix. Once the covariance matrix is obtained, the Mahalanobis distance can be calculated for a given observation. The formula for calculating the Mahalanobis distance is:

Mahalanobis Distance = $\sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$

x represents the vector of variables for the observation being evaluated. $\boldsymbol{\mu}$ represents the vector of means of the reference distribution. $-\sum(-1)$ represents the inverse of the covariance matrix of the reference distribution.

In simpler terms, the Mahalanobis distance is the Euclidean distance between the centered and scaled observation and the centered and scaled reference distribution, where the scaling is done using the covariance matrix [28].

4.1.3 Hotelling T2 test

The Hotelling T2 test, also known as Hotelling's T-squared test, is a multivariate statistical test used to determine if there is a significant difference between the means

Model	Mahalanobis Distance Mean	Standard deviation
Original	**71.6008**	**61.567**
GAN	43.598	28.625
GMM	**72.606**	**62.353**
VAE	59.472	43.776

Table 4.5: represents the Mahalanobis distance mean between the original dataset and the synthetic datasets.

of two groups in a multivariate setting. It is an extension of the univariate two-sample t-test to the multivariate case. In the multivariate setting, we have multiple variables or features measured on each observation. The Hotelling T2 test takes into account the covariance structure between these variables when comparing the means of two groups.

The test evaluates whether the mean vectors of the two groups are significantly different from each other in the multivariate space. It considers both the location and spread of the data in multiple dimensions. The test calculates a T2 statistic, which is a measure of the distance between the two mean vectors, taking into account the covariance structure [15][34][24]. The null hypothesis of the Hotelling T2 test is that there is no difference between the means of the two groups. The alternative hypothesis is that there is a significant difference between the means. The Hotelling T2 test requires certain assumptions to be valid, including multivariate normality and homogeneity of covariance matrices. Violations of these assumptions can affect the accuracy of the test results.

The test statistic follows Hotelling's T-squared distribution under the null hypothesis. Based on the calculated T2 statistic, along with the sample sizes and the number of variables, p-values can be obtained to determine the significance of the test. The Hotelling T2 test can also be applied to evaluate the quality and similarity of synthetic data compared to the original data. In the context of synthetic data evaluation, the Hotelling T2 test assesses whether the mean vectors of the synthetic data and the original data are significantly different from each other in a multivariate space. When evaluating synthetic data, the goal is to ensure that the synthetic

dataset captures the statistical properties and characteristics of the original dataset. The Hotelling T2 test can help determine if the synthetic data closely approximates the distribution of the original data by comparing their mean vectors.

The procedure for applying the Hotelling T2 test to synthetic data evaluation involves the following steps. Prepare the synthetic dataset and the original dataset, ensuring that they have the same number of variables or features. Calculate the mean vector for both the synthetic data and the original data by taking the average values of each variable across the respective datasets. Estimate the covariance matrices for both datasets. The covariance matrix describes the relationships and dependencies between the variables in the data. Compute the Hotelling T2 test statistic using the mean vectors and covariance matrices of the synthetic and original datasets. The test statistic quantifies the difference between the mean vectors, accounting for the covariance structure. Evaluate the significance of the test statistic by comparing it to the critical values from Hotelling's T-squared distribution. Calculate the p-value associated with the test statistic to determine whether the mean vectors of the synthetic and original datasets are significantly different. If the p-value is below a predetermined significance level (e.g., 0.05), it suggests that there is a significant difference between the synthetic and original datasets. On the other hand, a higher p-value indicates that the synthetic data closely resembles the original data in terms of their mean vectors. By applying the Hotelling T2 test, researchers and practitioners can quantitatively assess the quality and similarity of synthetic data to the original data. This evaluation method provides an objective measure of the extent to which the synthetic data captures the statistical characteristics of the original data.

Model	T2 test p-value
GAN	0.012
GMM	**1.11e-16**
VAE	4.66e-11

Table 4.6: represents the Hotelling T2 test p-value between the original dataset and the synthetic dataset generated by the VAE model.

4.2 Privacy and Information Preservation

Privacy and information preservation are crucial considerations when evaluating synthetic data. The evaluation process should ensure that the generated synthetic data preserves the privacy of individuals and does not disclose sensitive or personally identifiable information. Information preservation is an essential aspect of synthetic data evaluation, ensuring that the synthesized data retains the important statistical and structural properties of the original dataset [3]. The goal is to generate synthetic data that captures the essential characteristics and patterns of the original data while protecting sensitive information. Here are some techniques and considerations for information preservation in synthetic data evaluation:

- Statistical Properties: The synthesized data should maintain key statistical properties of the original dataset, such as distributions, correlations, and summary statistics. This ensures that the synthetic data accurately represents the underlying patterns and relationships present in the original data.

- Feature Preservation: It is important to preserve the essential features of the original dataset in the synthetic data. Features that are highly relevant to the analysis or modeling tasks should be accurately represented to maintain the data's utility and usefulness.

- Structural Integrity: The synthetic data should preserve the structural integrity of the original dataset. This involves maintaining the overall data structure, including hierarchical relationships, dependencies, or any other structural characteristics that are important for the analysis.

- Data Completeness: The synthetic data should adequately cover the range of values and patterns present in the original dataset. It should accurately represent the data's diversity and variability to ensure comprehensive analysis and modeling.

By focusing on information preservation during synthetic data evaluation, researchers can ensure that the generated data maintains its integrity, relevance, and usefulness. This facilitates reliable analysis, modeling, and decision-making processes while protecting sensitive information and complying with privacy regulations.

4.2.1 Mutual Information Score (MI)

Mutual Information (MI) is a statistical measure used in synthetic data evaluation to assess the amount of information shared between two variables or datasets. It quantifies the degree of dependency or association between variables, indicating how much knowledge of one variable can provide insights into the other.

In the context of synthetic data evaluation, MI is used to compare the information content of the original dataset and the synthetic dataset. It helps to determine how well the synthetic data captures the information present in the original data. A higher MI score indicates a stronger relationship and greater similarity between the variables or datasets being compared [20].

The MI score is based on the concept of entropy, which measures the uncertainty or randomness of a variable. By calculating the entropy of the original data and the synthetic data separately and then comparing their joint entropy, the MI score can be derived. The MI score can range from 0 to a maximum value, where 0 indicates no relationship or information shared, and the maximum value represents a perfect relationship.

In synthetic data evaluation, the MI score can be used to assess the fidelity and information preservation of the synthetic data. A higher MI score suggests that the synthetic data closely resembles the original data in terms of the underlying patterns, dependencies, and information content. On the other hand, a lower MI score indicates a potential loss of information or deviation from the original data's characteristics.

Let's consider two variables X and Y, where X represents the original dataset and Y represents the synthetic dataset.

Calculate the individual entropies of X and Y:

Entropy of X : $H(X) = -\sum p(x) \log p(x)$, where p(x) is the probability distribution of X. Entropy of Y : $H(Y) = -\sum p(y) \log p(y)$, where p(y) is the probability distribution of Y. Calculate the joint entropy of X and Y:

Joint Entropy of X and Y: $H(X, Y) = -\sum\sum p(x, y) \log p(x, y)$ is the joint probability distribution of X and Y. Compute the MI score using the formula:

MI(X, Y) = H(X) + H(Y) - H(X, Y) The MI score represents the reduction in uncertainty or the amount of information gained about one variable by knowing the other variable. A higher MI score indicates a stronger relationship or more shared

information between X and Y.

It's important to note that the estimation of probabilities and the calculation of entropy and joint entropy may vary depending on the specific data and the chosen estimation method. Different techniques such as histogram-based estimation, kernel density estimation, or nearest-neighbor methods can be employed to estimate the probability distributions.

By comparing the MI score between the original dataset (X) and the synthetic dataset (Y), researchers can assess the similarity in terms of the shared information and evaluate the fidelity and information preservation of the synthetic data [20][21].

In this test we take the MI score difference ie. $|\text{MI(synthetic)} - \text{MI(Original)}|$. The absolute value derived from this expression denotes the level of information retention that is present between the original dataset and the synthetic dataset generated by the three different models The scores obtained for the respective models are as follows.

Model	MI Score abs	MI Score synthetic
Original	——	0.99
GAN	0.77	0.22
GMM	**0.109**	**0.881**
VAE	0.44	0.55

Table 4.7: Represents the MI Scores of each model, depicting the information retention capabilities.

4.2.2 Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric method used to estimate the probability density function (PDF) of a random variable based on a given set of observations. It is commonly employed in synthetic data evaluation to assess the distributional similarity between the synthetic dataset and the original dataset [14].

The basic idea behind KDE is to represent the PDF as a weighted sum of kernel functions centered at each observation. The kernel function is a smooth, symmetric, and non-negative function that determines the shape of the estimated density

The KDE process involves the following steps. The choice of kernel function influences the smoothness and accuracy of the estimated density. The Gaussian kernel is a popular choice due to its smoothness and mathematical properties. The bandwidth parameter controls the width of the kernel function and affects the smoothness of the estimated density. A small bandwidth leads to a more detailed density estimate but may overfit the data, while a large bandwidth results in a smoother estimate but may over smooth the data. Selecting an appropriate bandwidth is crucial to obtain an accurate density estimate. For each observation in the dataset, the kernel function is centered at that point, and the density contribution is calculated. The contributions from all observations are summed to obtain the overall density estimate. The resulting estimated density function represents an approximation of the underlying PDF. It can be evaluated at any point to obtain the density value.

In the context of synthetic data evaluation, KDE can be used to estimate the PDF of the original dataset and compare it with the PDF of the synthetic dataset. By comparing the two density estimates, researchers can assess the similarity in terms of shape, peaks, and overall distributional characteristics. If the synthetic data closely matches the original data's density, it indicates a higher level of fidelity and preserves the statistical properties of the original dataset.

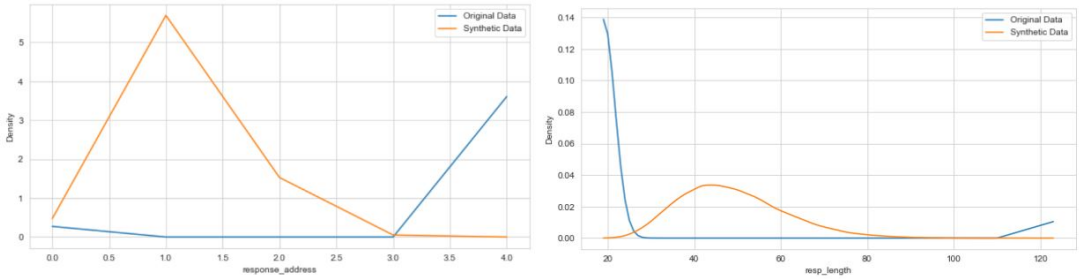

Figure 4.1: represents the KDE graph for GAN models, the comparison is made for the feature response_address (Left) and resp_length (right).

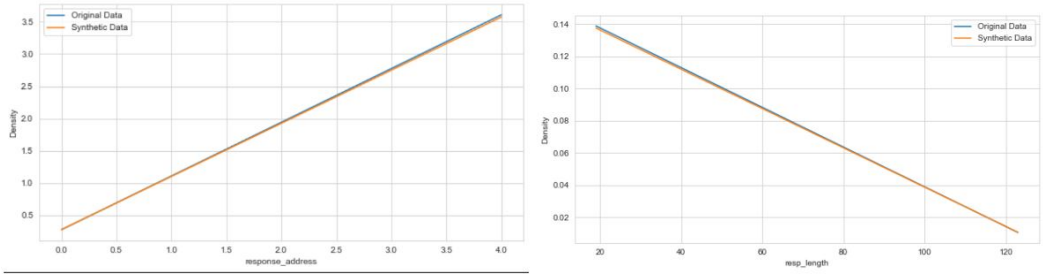

Figure 4.2: represents the KDE graph for GMM models, the comparison is made for the feature response_address (left) and resp_length (right).

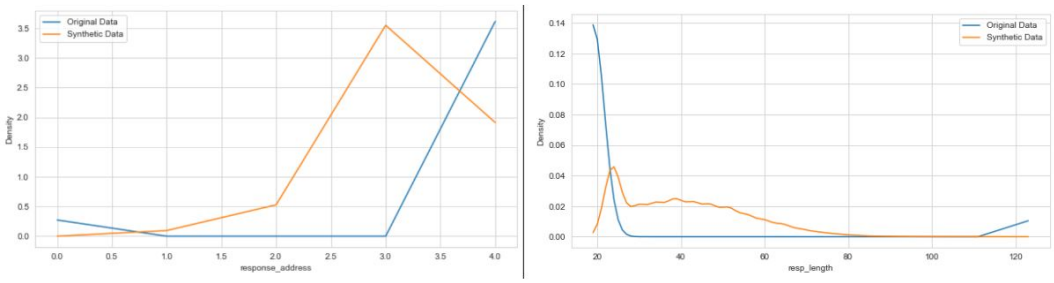

Figure 4.3: represents the KDE graph for VAE models, the comparison is made for the feature response_address and resp_length (right).

4.2.3 Wasserstein distance

The Wasserstein distance, also known as the Earth Mover's distance (EMD), is a metric used to quantify the difference between two probability distributions. It is often employed in synthetic data evaluation to assess the similarity between the probability distributions of the original and synthetic datasets [28][12].

Unlike other distance metrics that focus on comparing individual data points, the Wasserstein distance considers the entire distribution and measures the minimum cost of transforming one distribution into another. It accounts for the underlying structure and shape of the distributions, making it suitable for capturing differences in their probability density functions (PDFs).

The basic concept behind the Wasserstein distance involves considering the distributions as "piles of earth" and measuring the minimum amount of "work" required to transform one pile into the other. The work is calculated as the product of the amount of earth moved and the distance it is moved. This distance can be defined

based on various metrics, such as the Euclidean distance or the Minkowski distance.

However, computing the Wasserstein distance can be computationally expensive, especially for high-dimensional data or large datasets. Various algorithms, such as the Kantorovich-Rubinstein duality or Sinkhorn algorithm, have been developed to efficiently estimate the Wasserstein distance for practical applications.

In synthetic data evaluation, the Wasserstein distance can be used to compare the distributions of individual variables or the joint distribution of multiple variables between the original and synthetic datasets. By quantifying the dissimilarity, researchers can assess the quality and fidelity of the synthetic data generation process.

The mathematical working of the Wasserstein distance involves finding the optimal transportation plan that minimizes the cost of transforming one distribution into another. It considers the amount of mass being transported and the distance it needs to be moved. The distance metric used to measure the transportation cost can vary, but the most commonly used is the Euclidean distance.

Let's consider two probability distributions, P and Q, with probability density functions (PDFs) p(x) and q(x) respectively. The Wasserstein distance between P and Q, denoted as W(P, Q), can be calculated using the following formula:

$W(P, Q) = \inf \left(\int \int \|x - y\| \cdot T(x, y)\, dx\, dy \right)$

where the infimum is taken over all possible transportation plans T(x, y) that satisfy the following constraints:

Marginal constraints: The total amount of mass transported from each point in P must equal the corresponding point in Q. This ensures that no mass is lost or gained during the transportation process.

$$\int T(x, y) dy = p(x) \quad \text{for all } x$$

$$\int T(x, y) dx = q(y) \quad \text{for all } y$$

Non-negativity constraints: The transportation plan $T(x, y)$ should be non-negative, meaning that it cannot transport negative mass.

$$T(x, y) \geq 0 \quad \text{for all } x, y$$

The Wasserstein distance measures the minimum cost of transporting mass from P to Q, where the cost is determined by the distance $\|x-y\|$ multiplied by the amount of mass being transported $T(x, y)$. By finding the optimal transportation plan that minimizes this cost, the Wasserstein distance provides a measure of dissimilarity between the two distributions [28][2].

Using this as the base the Wasserstein distance is measured between the original dataset and the synthetic dataset. The results obtained are as follows: measure the distance between two probability distributions

A low Wasserstein distance between the original and synthetic data indicates that the synthetic data captures the same distribution as the original data.

Model	Wasserstein distance
GAN	0.534
GMM	**0.106**
VAE	0.371

Table 4.8: Represents the Wasserstein distance Scores of each model

4.2.4 t-SNE plot visualizations

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique commonly used for visualizing high-dimensional data in a lower-dimensional space. It is often employed in synthetic data evaluation to examine the structure and clustering patterns of the synthetic data in comparison to the original data [24].

The working of t-SNE involves transforming the high-dimensional data into a two- or three-dimensional representation while preserving the local similarities between data points. It achieves this by modeling pairwise similarities using probability distributions. The algorithm starts by computing pairwise similarities between the high-dimensional data points using a Gaussian kernel. The similarities are then converted into probabilities using a Student's t-distribution.

The t-SNE algorithm iteratively maps the high-dimensional data points to the low-dimensional space, aiming to minimize the divergence between the pairwise similarity distributions of the high-dimensional data and the low-dimensional embeddings. The

mapping process is performed by optimizing the Kullback-Leibler (KL) divergence between the two distributions.

This is a powerful visualization technique that enables the exploration and comparison of high-dimensional data in a lower-dimensional space. It provides a visual representation of the structures and clustering patterns of the data. This can assist in determining the similarity and quality of synthetic data when compared with the original data [24].

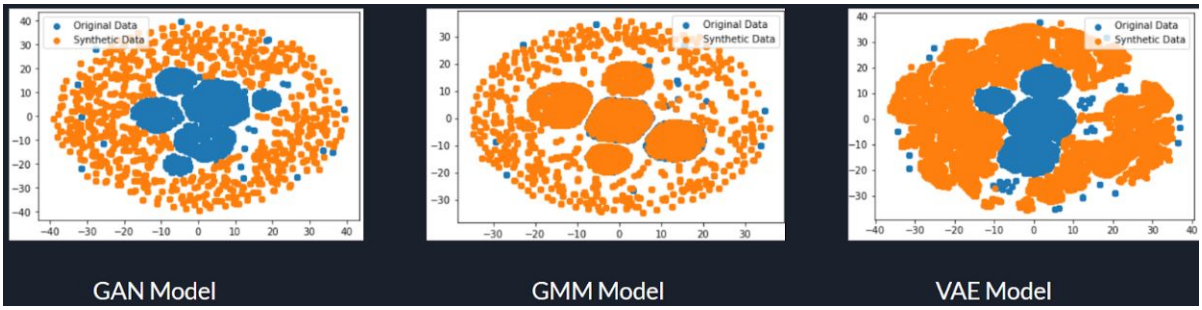

Figure 4.4: Represents the t-SNE plot visualizations that are performed in order to see if the original dataset and the synthetic dataset have similar clusters.

4.3 Diversity and Generalization

Evaluating the diversity of synthetic data is an essential aspect of assessing the quality and usefulness of generated synthetic datasets. Diversity refers to the extent to which the synthetic data captures the variability and distribution of the original data. A diverse synthetic dataset should accurately represent the different patterns, relationships, and characteristics present in the original dataset [28].

Evaluating diversity is crucial because a lack of diversity in synthetic data can lead to biased or incomplete representations, which may limit the effectiveness and applicability of the generated data. Therefore, it is important to employ evaluation measures and techniques that quantitatively assess the diversity of the synthetic data.

There are several approaches to evaluating the diversity of synthetic data. One common method is to compare statistical properties between the original and synthetic datasets. This can include analyzing summary statistics, distributional characteristics, or correlations of relevant variables. If the synthetic data closely matches these statistical properties, it indicates a higher level of Diversity [37].

Additionally, visualization techniques can be employed to visually inspect the diversity of the synthetic data. Scatter plots, histograms, or t-SNE plots can help visualize the distribution and clustering patterns of the synthetic data in comparison to the original data. Visual assessments can provide valuable insights into the diversity of the synthetic data and help identify any discrepancies or limitations.

4.3.1 Histograms and Scatter Plots

In order to compare and visualize the distributions of particular features in the original and synthetic datasets Histograms prove to be one of the most useful tools. Taking a particular feature into consideration that belongs to both the original dataset and the synthetic dataset when they are plotted, it becomes easier to identify and analyze discrepancies or similarities between the two datasets.

Using this method, if the features of the synthetic dataset closely match that of the original dataset, then it is likely that the synthetic dataset has a good representation of the original data. Likewise, if there are a lot of discrepancies between the two features on the histogram then the synthetic data does not capture the distribution of the features correctly [37].

Histograms can also be used to identify any outliers in the synthetic data. If there are large spikes or gaps in the histogram of the synthetic data, it could indicate the presence of outliers that need to be investigated further.

Overall, histograms provide a quick and easy way to compare the distribution of a feature between the original and synthetic data and can be a useful tool in synthetic data analytics.

Similar to the histograms scatterplots are a useful visualization tool for synthetic data analytics. They are used to plot the relationship between two variables in a dataset. In the context of synthetic data analytics, scatterplots are used to visually compare the distribution of variables in the original and synthetic datasets [36].

If the scatterplot shows a linear relationship between the variables in both datasets, it suggests that the synthetic data is a good representation of the original data. However, if the scatterplot shows a nonlinear relationship or there are significant differences between the variables in the two datasets, it may indicate that the synthetic data needs further refinement.

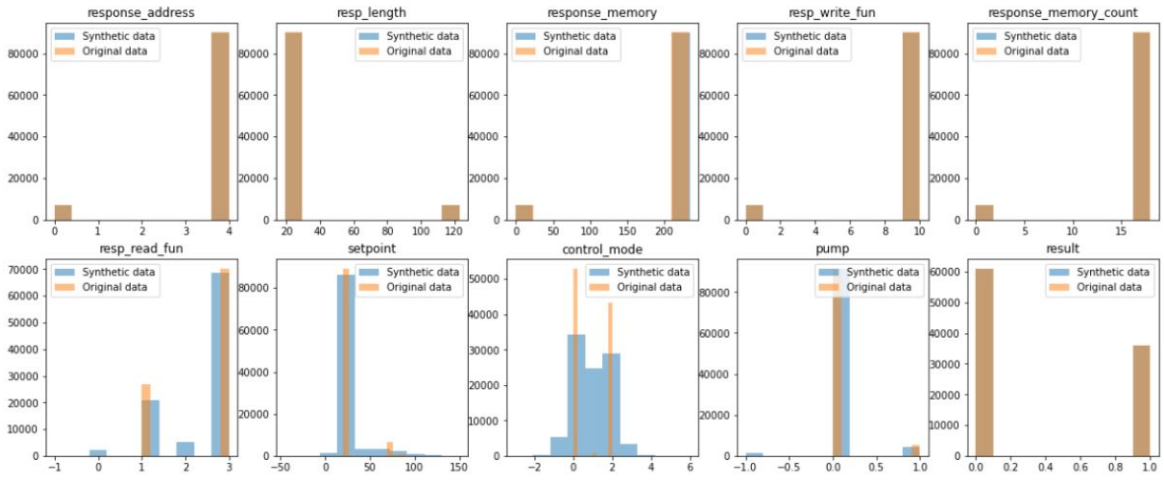

Figure 4.5: Represents the different feature distributions in the form of Histograms for original data compared with synthetic data generated by the GMM model.

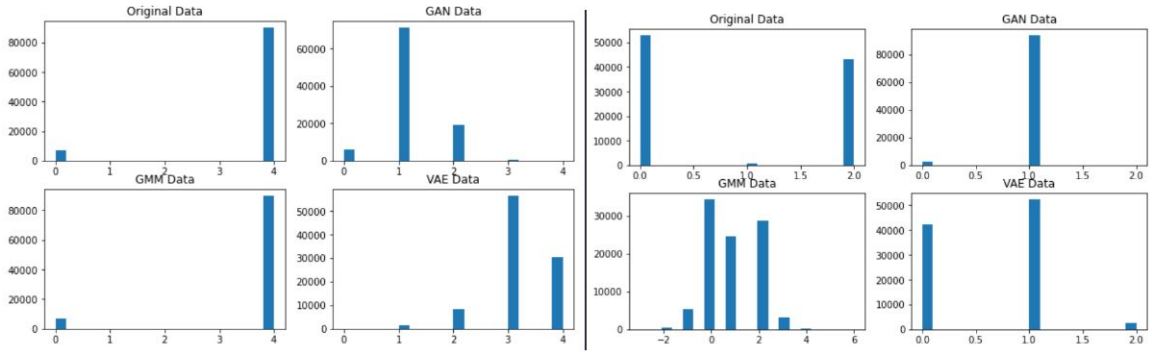

Figure 4.6: Represents the distribution of a particular feature compared to the original data and the synthetic data generated by the three different models GAN, GMM, and VAE in the form of Histograms.

In addition to comparing the distribution of variables, scatterplots can also be used to identify outliers in the synthetic dataset. Outliers are data points that deviate significantly from the overall pattern of the data. If the scatterplot shows outliers in the synthetic dataset that were not present in the original data, it may indicate that the synthetic data is not representative of the original data and needs to be refined [31].

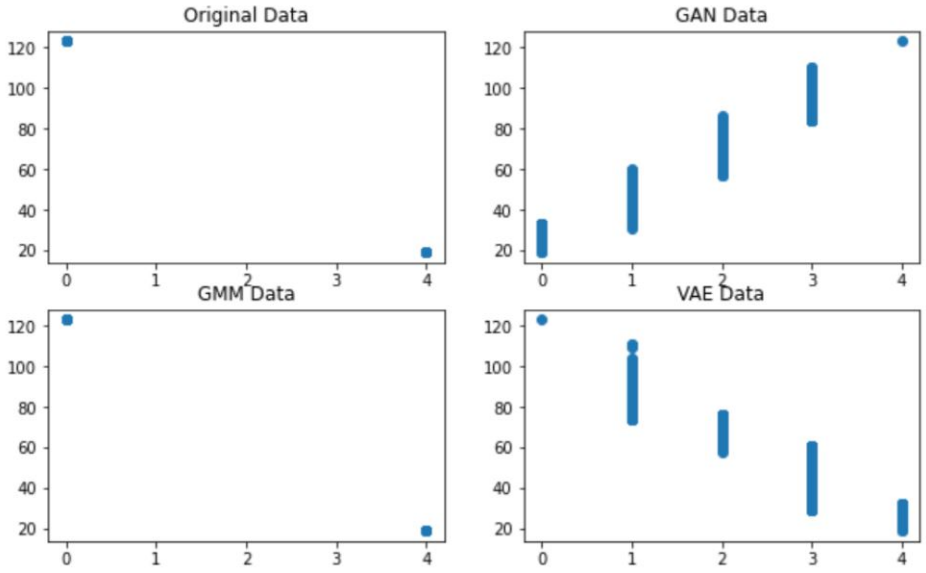

Figure 4.7: Represents the distribution of a particular feature compared to the original data and the synthetic data generated by the three different models GAN, GMM, and VAE in the form of scatter plots.

4.3.2 Kolmogorov-Smirnov (KS) test

The Kolmogorov-Smirnov test is a statistical test used to evaluate the similarity between two probability distributions. In the context of synthetic data evaluation, the Kolmogorov-Smirnov test can be employed to assess the similarity between the distribution of the original dataset and the distribution of the synthetic dataset [20].

The test works by comparing the empirical cumulative distribution functions (CDFs) of the two datasets. The empirical CDF represents the proportion of data points in a dataset that are less than or equal to a given value. The Kolmogorov-Smirnov test calculates the maximum absolute difference (D) between the two empirical CDFs. This difference, also known as the Kolmogorov-Smirnov statistic, serves as a measure of dissimilarity between the distributions.

To perform the Kolmogorov-Smirnov test, the null hypothesis is that the two datasets are drawn from the same distribution. If the calculated Kolmogorov-Smirnov statistic exceeds a critical value (typically obtained from statistical tables), the null hypothesis is rejected, indicating that the two distributions significantly differ from each other [23].

In the context of synthetic data evaluation, the Kolmogorov-Smirnov test can be

used to determine whether the distribution of the synthetic dataset matches that of the original dataset. A smaller Kolmogorov-Smirnov statistic suggests a higher similarity between the distributions, indicating that the synthetic data is more representative of the original data. The results obtained ie. the corresponding p-value to each synthetic data generation model is as follows:

Model	Kolmogorov-Smirnov test p-value
GAN	0.12
GMM	**3.1175 e-29**
VAE	0.000521

Table 4.9: Represents the Kolmogorov-Smirnov (KS) test Scores of each model

4.4 Interpretability and Utility

Evaluating the model interpretability and utility of synthetic data is crucial to assess the effectiveness and practicality of using synthetic data in various applications. Model interpretability refers to the ability to understand and interpret the inner workings and decisions made by a model, while utility refers to the usefulness and effectiveness of the synthetic data in achieving the intended goals [33][21].

In the context of synthetic data evaluation, model interpretability involves assessing how well the synthetic data captures the underlying patterns, relationships, and features present in the original data. It requires evaluating whether the synthetic data retains the important characteristics of the original data, such as the distributional properties, correlations between variables, and relevant patterns or trends. Various techniques can be employed to assess interpretability, including visualizations, statistical analysis, and comparison with domain knowledge.

On the other hand, evaluating the utility of synthetic data focuses on determining the extent to which the synthetic data can effectively substitute the original data in practical applications or analyses. This involves evaluating the performance of models or algorithms trained on the synthetic data and assessing their ability to achieve the desired outcomes. Utility evaluation may involve comparing the performance of models trained on the synthetic data with models trained on the original data, considering metrics such as accuracy, precision, recall, or any other relevant performance

measure [18].

4.4.1 Decision Trees and Feature Importance

Evaluating the model interpretability of synthetic data is essential to understand the underlying patterns and decision-making processes of the generated data. One popular approach for assessing interpretability is through the use of decision trees.

Decision trees are hierarchical models that make decisions based on a sequence of rules or conditions. They provide a transparent and intuitive representation of how the input variables influence the output or target variable. When applied to synthetic data evaluation, decision trees can reveal the important features and relationships within the data, helping to understand how the synthetic data captures the patterns present in the original data [6].

The use of decision trees for the model interpretability of synthetic data offers several advantages. Decision trees provide a clear and transparent representation of the decision-making process, making it easier to understand and interpret the generated data. They also allow for the identification of important features and their relative importance, aiding in feature selection and dimensionality reduction. Furthermore, decision trees can handle both numerical and categorical variables, making them suitable for various types of data.

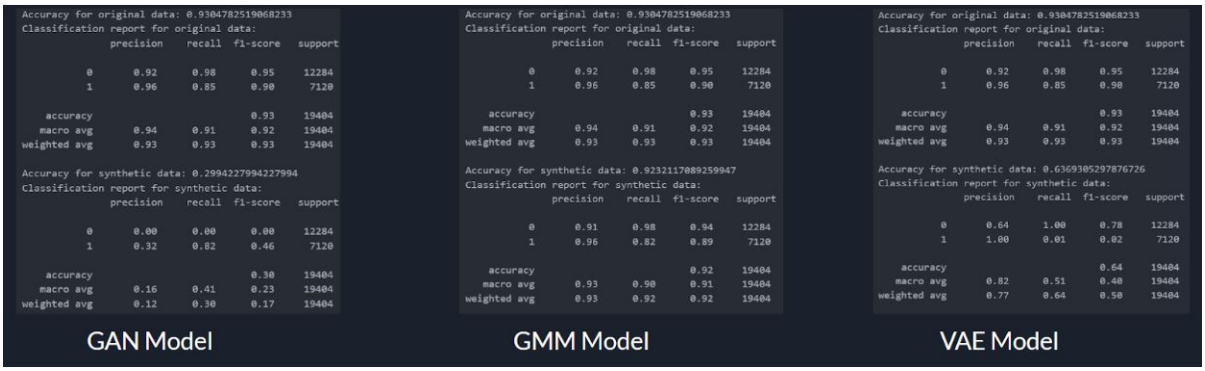

Figure 4.8: Represents the performance of each synthetic data generated by the respective model's GAN, GMM, and VAE models against the original data using decision trees.

The second section of this test is performed to check if the original dataset and the synthetic data are from the same distribution or not. This is done by merging the

two datasets and adding a target column to indicate if the corresponding instance is from the original dataset or the synthetic dataset. The model's capability to detect if the instance is real or synthetic can determine if the dataset has been spliced from one another ie. the lower the prediction accuracy greater the chance of datasets being generated from different distributions [33]. The results are as follows:

Model	Accuracy of the Decision tree
GAN	72%
GMM	**46%**
VAE	58%

Table 4.10: Represents the prediction accuracy rate of each model

The third section of this test is performed in order to determine the feature importance of each feature in the original dataset and the synthetic dataset and compare them. Feature importance analysis provides insights into the relevance and contribution of individual input variables in the model's decision-making process [33].

The process of evaluating model interpretability using feature importance typically involves the following steps. The synthetic data is used to train a machine learning model, such as a decision tree, random forest, or gradient boosting model. These models are capable of capturing complex relationships between the input variables and the target variable. After training the model, the feature importance is calculated based on the model's internal mechanisms. Various techniques can be employed to measure feature importance. In linear models or models with interpretable coefficients (e.g., logistic regression), the magnitude of the coefficients can be used to assess feature importance. Larger coefficients indicate greater importance.

Evaluating model interpretability using feature importance provides several benefits. It allows researchers to understand the factors influencing the model's predictions and identify the most influential features in the synthetic data. This knowledge can help detect potential biases, assess the generalizability of the synthetic data, and guide further data generation or refinement efforts [33][21].

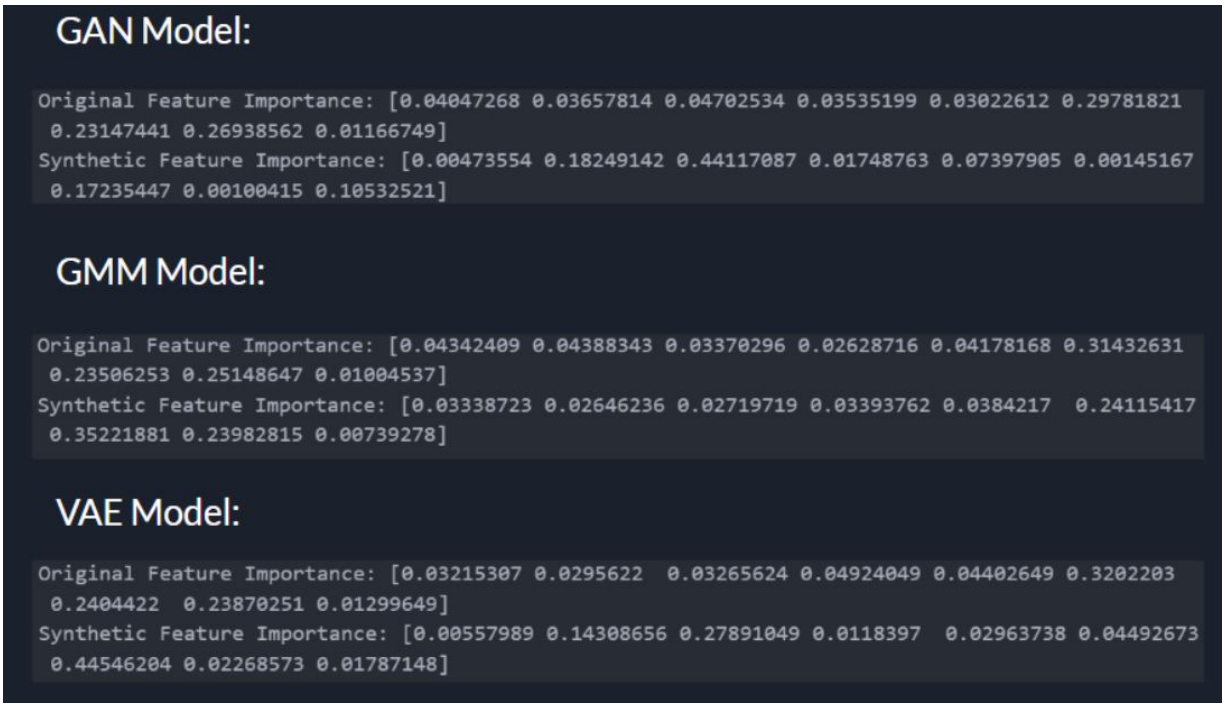

Figure 4.9: Feature importance with reference to model coefficients.

4.5 Comparative Analysis

Using the Results obtained in the above section, a comparative analysis report is formed. This section of the thesis contains a comprehensive analysis of the results obtained above and revisiting and explaining the significance of each subsection. The evaluation is divided into four subsections as per the suggested evaluation framework under the subheadings of Fidelity, Privacy and Information Preservation, Diversity and Generalization, and finally Interpretability and Utility.

Through the above subsections, each topic is explored along with the corresponding test being either statistics, visualization, or machine learning oriented. In this section, we aim to draw inferences with respect to the results obtained.

To begin with fidelity, initially, a descriptive statistics methodology is employed to provide a detailed statistical analysis of the various elements the synthetic data that has been generated is to follow. The results obtained from each dataset are displayed and compared.

Using the Discriptive Statistics alone it is evident that the GMM model generated the synthetic data that is most alike to the original dataset. This test mainly handles

Type of Dataset	Parameter	response_address	resp_length	response_memory	resp_write_fun	response_memory_count	resp_read_fun	setpoint	control_mode	pump	result
Original data	Count	97019	97019	97019	97019	97019	97019	97019	97019	97019	97019
GAN data											
GMM data											
VAE data											
Original data	**Mean**	**3.719**	**26.29**	**216.65**	**9.29**	**16.73**	**2.44**	**24.16**	**0.899**	**0.056**	**0.36**
GAN data		1.14	48.77	66.48	2.88	5.15	1.98	39.77	0.97	0.953	0.053
GMM data		**3.714**	**26.29**	**216.4**	**9.286**	**16.71**	**2.34**	**24.26**	**0.899**	**0.033**	**0.37**
VAE data		3.196	40.73	185.2	7.621	14.24	2.649	34.318	0.589	0.047	0.63
Original data	**Std**	**1.021**	**26.56**	**59.5**	**2.55**	**4.59**	**0.893**	**14.32**	**0.991**	**0.23**	**0.482**
GAN data		0.512	12.12	27.07	1.19	2.11	0.193	8.101	0.17	0.211	0.224
GMM data		**1.029**	**26.77**	**59.99**	**2.574**	**4.63**	**0.903**	**14.509**	**1.026**	**0.238**	**0.483**
VAE data		0.652	14.38	32.32	1.382	2.565	0.490	9.645	0.539	0.211	0.2249

Table 4.11: indicates the accumulate of the descriptive statistics for the different datasets

the primary statistical features that the model is trained to replicate. The Synthetic Data Generation Model is expected to generate data that is capable of substituting the original data, thus the primary function of replicating or producing results close to that of the original data is paramount. Supplementing the results of the descriptive statistics methodology tests to provide additional information on the fidelity of the synthetic data are depicted in Tab.13.

Model	Mahalanobis Distance Mean	Standard deviation	T^2 test p-value
Original	**71.6008**	**61.567**	———————
GAN	43.598	28.625	0.012
GMM	**72.606**	**62.353**	**1.11e-16**
VAE	59.472	43.776	4.66e-11

Table 4.12: indicates the accumulated results of the Mahalanobis Distance and the T2 test's p-value

The Mahalanobis Distance depicts the overall distance between two datasets, as per the tests definition the closer the mean distance to the original data the more significant the similarity. And to supplement the result of this test, a T2 test is also performed where the p-value is taken into consideration. The null hypothesis taken into consideration when performing the T2 test is that the two datasets are from the same distribution for p-value approaches 1. Thus using the scores of this test it can be seen that the GMM model possesses a p-value that is closest to 0, followed by the VAE model and then the GAN Model. This pattern is similarly replicated in the results of the Mahalanobis Distance measurements as well.

Next, the evaluation procedure that analyses Privacy and information Preservation is performed using the help of the MI score and the Wasserstein Distance. The

MI Score in particular is useful when it comes to comparing the information content of the original dataset and the synthetic dataset. It helps to determine how well the synthetic data captures the information present in the original data. A higher MI score indicates a stronger relationship and greater similarity between the variables or datasets being compared. The Wasserstein Distance is often employed in synthetic data evaluation to assess the similarity between the probability distributions of the original and synthetic datasets. The MI score and the Wasserstein Distance show a similar performance trend when it comes to the synthetic datasets. In these tests, the results on Tab 4.12 are the absolute differences between the scores for the original and the synthetic dataset. The GMM model possesses the greatest information preservation capabilities and also displays the greatest similarity when it comes to the distribution of the dataset. This information preservation capability is further explored when considering the KDE plots that can determine how the synthetic data is distributed when compared with the original data. The results of the KDE test can be seen in Fig 4.1,4.2 and 4.3. The distribution characteristics shown by the GMM model are very similar to the original data, this could mean that the data has fulfilled one of the main purposes of generating synthetic data ie. to substitute the original data. The VAE model generates synthetic data that follows a similar pattern to that of the original model where it imitates the peaks and lulls but also provides a diverse take on the dataset, meaning that the dataset contains a variety of use case scenarios that would be present in a controlled environment such as the original dataset. On the other hand, the GAN models result is not as what is expected as it produces a diverse dataset that does not have a lot of similarity with the original dataset.

Model	MI Score Difference	Wasserstein distance Difference
GAN	0.77	0.534
GMM	**0.109**	**0.106**
VAE	0.44	0.371

Table 4.13: indicates the accumulated results of the MI Score and Wasserstein Distance.

Continuing on with Diversity and Generalization, this is explored in depth with the assistance of Histogram plots, Scatter plots, and the KDE plot from the previous

section Fig 4.1-3, Fig 4.5-7. The results obtained from these plots display a continuing pattern of GMM imitating the original dataset with the greatest likeliness, with the VAE model displaying variety and diversity but also maintaining a similar pattern, while the GAN models fall behind these models in this regard. To provide conclusive evidence of the above result, a Kolmogorov-Smirnov test is also performed. The Kolmogorov-Smirnov test can be used to determine whether the distribution of the synthetic dataset matches that of the original dataset. A smaller Kolmogorov-Smirnov statistic suggests a higher similarity between the distributions, indicating that the synthetic data is more representative of the original data. The results of this test can be seen in Tab 4.9.

The need for diversity is a growing need in the current day society where ICS datasets are of immense value and use. Real-time data lacks adverse case scenarios, where the ICS detects the signals present during the other end of the spectrum. This case scenario is least explored and not anticipated as much, therefore in accordance with the original data, the VAE Model would perform the best with regards to diversity while also having a good performance when it comes to predictive analysis. This can be seen in the previous subsection where interpretability and utility are discussed.

In this case, the experiments performed are by training the model on a predictive algorithm to determine its usability and performance. The models that have been trained on a decision tree are evaluated based on their predictive performance. Another version of the test is also performed to test the predictive capability of the model when trained on a dataset that possesses the combined result of the original dataset and the synthetic dataset of the respective model. The results proved conclusive as the GMM model performs the best with very high accuracy and performance, followed by the VAE model and GAN model respectively which could be seen in Fig.13. The second test also proved insightful as could be seen in Tab.11. The next test performed is to analyze the feature importance and if the weights assigned to each feature per model is similar or not. This can be done using a Python extension and the results can be seen in Fig.14. The feature importance values although not the same in nature, still do follow a similar pattern as above.

4.6 Discussions and Remarks

Different methods of generating high quality synthetic data for Industrial Control Systems (ICS) (GAN, VAE and the GMM models) was evaluated using visual analytics and statistical techniques. This research aimed on addressing a few factors such as the need for realistic and diverse datasets, datasets that have high levels of information preservation, and a suitable framework for evaluating the generated synthetic data. This thesis used a comprehensive comparative analysis to determine the best method.

An evaluation framework allowing us to quantify the similarity between synthetic and real datasets, measure data fidelity, evaluate information preservation, and assess the diversity and representativeness of the synthetic data was developed. We gained a deeper understanding of the characteristics of the synthetic data and made informed decisions regarding its suitability for different research and application contexts. The evaluation framework developed in this research facilitated the assessment of the synthetic data quality from various perspectives. The fields that have the highest need for synthetic data are evaluated based on their preference. Fidelity measures such as the Mahalanobis distance and the T2 test allowed for quantification of similarity between the original dataset and the synthetic dataset. Information preservation metrics such as the MI score and the KDE plots provided an accurate representation of the dataset's capability to preserve the essential information in the synthetic data. t-SNE plot visualizations aids in evaluating their interpretability and utility with machine learning techniques.

Through our comparative analysis, we evaluated three prominent techniques for generating synthetic data: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Gaussian Mixture Model (GMM). Each technique was assessed in terms of its effectiveness, performance, and applicability to ICS applications. Our findings indicate that the Gaussian Mixture Model (GMM) outperformed the other techniques in terms of fidelity and performance. It demonstrated the highest ability to capture the statistical properties and patterns of real-world ICS datasets, making it the ideal choice for generating high-quality synthetic data in ICS.

One of the major research goals this thesis set out to achieve was to identify a synthetic data generation model that could aid researchers overcome the limitations

of privacy and provide synthetic data that is identical to the original dataset. From the comparative analysis it can be seen that the GMM model excels in replicating and generating synthetic data that is extremely close to the original dataset over its competitors. The GMM model demonstrated a strong performance in capturing the statistical analysis properties of the original data, thus making it the ideal candidate in order to fulfil the primary study objective. The research also identified research gaps and challenges in the domain of generation and evaluation of synthetic data for ICS datasets and the need for further exploration of privacy and such information preservation techniques, capabilities in capturing dependencies and also addressing concerns regarding utility and scalability.

Information preservation is a key component of synthetic data generation. It is important for the synthetic data to preserve the characteristics as much possible from the original dataset. However, it is also important to have the capability to manipulate the synthetic data generation in such a way that it also involves improving adverse case scenarios to the dataset in question. This will enable to decrease the bias involved in prediction or analysis of synthetic data. In this case the VAE model excels in preserving the data and as well as generating a diverse dataset followed by the GAN models. Thus, this thesis provides valuable insights and aims to be a foundation to generate high-quality synthetic data for ICS datasets and provide an evaluation framework that is freely customizable to the user's everyday needs and also contributes to the advancements in these areas.

Chapter 5

Conclusion

In conclusion, through the comparative analysis of the three models the GAN, VAE and the GMM models, it was observed that the CTGAN model performs well in enabling fine-grained control over the data generation. However, it is seen that the model requires further fine tuning to be a suitable substitute to the original dataset or to create highly diverse datasets that could serve as an unbiased class of data. Outcomes of this thesis contribute to the advancement of the field of ICS research along with providing valuable insights into the strengths and weaknesses of the generation of synthetic data. Researchers can leverage this knowledge to select the right methods based on their needs and application processes. The availability of such high-quality datasets will assist people in future endeavours as high-quality sensor datasets will be readily available for research purposes and not be inhibited due to privacy concerns. This would also enable the development of algorithms tailored for ICS datasets and also improve the system designs and enhanced security mechanisms of the ICS.

5.1 Future Work

It is worth noting that this research is not without limitations. While our evaluation framework provided a comprehensive analysis, there may be additional factors or metrics that could be considered in future studies. Furthermore, the choice of synthetic data generation technique may vary depending on the specific characteristics of the ICS dataset and the research objectives.

It is also recommended to work on other forms of synthetic datasets and also explore Transformer based models and other such hybrid models to diversify the evaluation result. Continual evaluation of the ICS datasets and improvement of the framework is paramount for success to address the future challenges experienced in the various synthetic data generation models.

Thus, this thesis provides valuable insights and aims to be a solid foundation

to generate high-quality synthetic data for ICS datasets and provide an evaluation framework that is freely customizable to the user's everyday needs and also contributes to the advancements in these areas.

Bibliography

[1] T. Akhtar and BB Gupta. Analysing smart power grid against different cyber attacks on scada system. *International Journal of Innovative Computing and Applications*, 12(4):195–205, 2021.

[2] A. Alshehri, R. Mahapatra, and K. McLaughlin. Synthetic ics sensor data generation using gans and lstm networks. *IEEE Transactions on Industrial Informatics*, 16(4):2563–2573, 2020.

[3] D. L. Banks, P. D. Ruggiero, and A. B. Ruggiero. Synthetic data in computational statistics. In *Handbook of Computational Statistics*, pages 931–951. Springer, 2004.

[4] J. S. Breese and D. Heckerman. Decision-theoretic debugging. In *Proceedings of the 13th conference on Uncertainty in Artificial Intelligence*, pages 55–64, 1997.

[5] J. Chen, C. Yang, X. Huang, C. Xiao, and G. Yang. Differentially private synthetic data generation for industrial control systems. *IEEE Transactions on Industrial Informatics*, 17(3):2019–2029, 2021.

[6] Lei Chen, Yuan Li, Xingye Deng, Zhaohua Liu, Mingyang Lv, and Hongqiang Zhang. Dual auto-encoder gan-based anomaly detection for industrial control system. *Applied Sciences*, 12(10):4986, 2022.

[7] P. Chen, Z. Yu, and Y. Xiao. Synthetic data generation for industrial control systems using bayesian networks. *IEEE Transactions on Industrial Informatics*, 13(6):3124–3134, 2017.

[8] K. S. Deepak, M. Chakraborty, and R. Jana. Cyber-physical system security: A case study on the stuxnet worm. In *2011 Annual IEEE India Conference*, pages 1–6. IEEE, 2011.

[9] D.Upadhyay and S.Sampalli. Scada (supervisory control and data acquisition) systems: Vulnerability assessment and security recommendations. *Computers and Security*, 89:101666, 2020.

[10] Khaled E.Emam. Seven ways to evaluate the utility of synthetic data. *IEEE Security & Privacy*, 18(4):56–59, 2020.

[11] D. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[12] J. Giraldo et al. Evaluation of synthetic power system data considering the system behavior. *IEEE Transactions on Smart Grid*, 10(6):6267–6276, 2019.

[13] Andre Goncalves et al. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(1):1–40, 2020.

[14] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability, and Security*, 2019.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[16] E. Hodo et al. Metrics for evaluating the quality of synthetic intrusion detection system datasets. *Computers & Security*, 68:18–31, 2017.

[17] A. E. Howe, H. Fei, J. S. Pupillo, and A. J. Medford. Synthesized measurement data in support of nuclear safeguards decision making. In *International Topical Meeting on Nuclear Safeguards and Security*, 2013.

[18] Y. Li, Z. Tang, X. Zhu, and F. Wu. Generating synthetic datasets for anomaly detection in industrial control systems. In *Proceedings of the 15th IEEE International Conference on Networking, Sensing, and Control*, pages 1–6, 2018.

[19] J. Liu, H. Gao, H. Li, Y. Song, and W. Han. Synthetic data generation using deep learning for anomaly detection in industrial control systems. *Applied Sciences*, 10(10):3429, 2020.

[20] X. Liu et al. Evaluating the utility of synthetic data for differentially private time series classification. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1503–1516, 2020.

[21] E. Monteiro et al. A visual analytics approach for synthetic ics data evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):979–988, 2020.

[22] Mohammad Noorizadeh, Mohammad Shakerpour, Nader Meskin, Devrim Unal, and Khashayar Khorasani. A cyber-security methodology for a cyber-physical industrial control system testbed. *IEEE Access*, 9:16239–16253, 2021.

[23] L. Qian, X. Zhu, Y. Li, and F. Wu. Synthetic data generation for anomaly detection in industrial control systems using generative adversarial networks. *IEEE Transactions on Industrial Informatics*, 15(1):487–495, 2018.

[24] A. Salem et al. Evaluating the privacy and utility of synthetic social network data. *ACM Transactions on Privacy and Security*, 21(2):1–28, 2018.

[25] S. Song et al. Statistical evaluation of synthetic power system data. *IEEE Transactions on Power Systems*, 33(3):2817–2825, 2022.

[26] Christopher T Symons and Justin M Beaver. Nonparametric semi-supervised learning for network intrusion detection: combining performance improvements with realistic in-situ training. In *Proceedings of the 5th ACM workshop on Security and artificial intelligence*, pages 49–58, 2012.

[27] J. Wang, X. Liu, and Y. Liu. Synthetic data generation for industrial control systems using variational autoencoders. *Sensors*, 20(21):6247, 2020.

[28] W. K. Wong et al. Visual data comparison and its application in evaluating synthetic data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):369–378, 2016.

[29] Y. Wu, C. Xu, K. Zhang, and X. Hu. Visualization techniques for evaluating synthetic industrial control system datasets. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):949–958, 2019.

[30] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*, 2018.

[33] SLP Yasakethu and J Jiang. Intrusion detection via machine learning for scada system protection. In *1st International Symposium for ICS & SCADA Cyber Security Research 2013 (ICS-CSR 2013) 1*, pages 101–105, 2013.

[34] B. Zhang, X. Liu, and J. Wang. Synthetic data generation for industrial control systems using generative adversarial networks. *IEEE Access*, 8:12370–12380, 2020.

[35] K. Zhang, Y. Wu, and X. Hu. Synthetic ics datasets generation framework for training intrusion detection systems. *IEEE Transactions on Industrial Informatics*, 15(11):6013–6022, 2019.

[36] M. Zhang, Z. Ren, Y. Xiang, and S. Mao. A statistical framework for evaluating synthetic datasets for industrial control systems. *IEEE Transactions on Industrial Informatics*, 16(7):4450–4461, 2020.

[37] Y. Zhang et al. Evaluating the utility of generative adversarial networks in synthetic medical data generation. *Artificial Intelligence in Medicine*, 82:45–56, 2016.

[38] H. Zou, L. Huang, and H. H. Ali. Generative adversarial networks-based synthetic data generation for industrial control systems. *IEEE Access*, 7:45142–45152, 2019.