Enhancing Data Security in Healthcare with Synthetic Data Generation: An Autoencoder and Variational Autoencoder Approach

Kelechukwu Innocent Ede

Thesis submitted for the degree of
Master in Applied Computer and Information Technology - ACIT
(Data Science)
30 credits

Department of Computer Science
Faculty of Technology, Art and Design

Oslo Metropolitan University — OsloMet

Spring 2024

Enhancing Data Security in Healthcare with Synthetic Data Generation: An Autoencoder and Variational Autoencoder Approach

Kelechukwu Innocent Ede

Enhancing Data Security in Healthcare with Synthetic Data Generation: An Autoencoder and Variational Autoencoder Approach

http://www.oslomet.no/

Printed: Oslo Metropolitan University — OsloMet

Abstract

The advent of machine learning and artificial intelligence (AI) in healthcare has revolutionized data analysis and patient care. However, utilizing real patient data presents substantial privacy and security challenges. This thesis tackles these challenges by exploring the application of Auto-Encoders (AEs) and Variational Auto-Encoders (VAEs) in synthetic healthcare data generation, offering an alternative to the typical use within Generative Adversarial Networks (GANs).

While techniques like CTGAN are known for generating realistic synthetic data, there is variability in how different implementations address the security of the original data during generation. This study leverages the unique data encoding capabilities of AEs and VAEs to propose a method that enhances data security, thereby producing synthetic data that upholds privacy while retaining utility for AI applications in healthcare, such as disease diagnosis and predictive modeling.

The methodology was rigorously tested across three diverse healthcare datasets, varying in size and characteristics, to ensure the effectiveness of the proposed solutions in protecting original data privacy while generating high-quality synthetic data. These methods were further evaluated using the Anonymeter tool to assess privacy risks thoroughly, ensuring a robust validation against the datasets used in prior research and affirming the advancements made by integrating AEs and VAEs.

This work contributes to the field of healthcare AI by providing a secure data generation framework that balances data utility with privacy. It sets the stage for future research in developing privacy-compliant AI systems in healthcare, highlighting the potential for widespread application of synthetic data while maintaining stringent privacy standards.

For more information, detailed implementations, and additional resources, please visit the project's: GitHub Repository.

Keywords: AI, Auto-Encoders, Variational Auto-Encoders, synthetic data, data privacy, Anonymeter, healthcare datasets, machine learning, predictive modeling.

Acknowledgments

I would first like to give all thanks and glory to God for His blessings, good health, and continuous guidance. I am immensely grateful to my project supervisors, Hårek Haugerud and Anis Yazidi, for their expert guidance and unwavering support, which were pivotal to my research. Their invaluable advice and insightful feedback have profoundly shaped both this project and my academic growth. I also extend my deepest appreciation to my parents for their love and sacrifices, my wife, Ifeyinwa Juliana Ede, and my mother, Ogechukwu Elizabeth Ede, for their endless support and patience, and my friends for their encouragement and fellowship throughout this journey.

Each of you has played a crucial role in this project, and I am sincerely thankful for the strength and inspiration you have provided me. Your collective support has been my cornerstone, making this journey not only possible but also a truly enriching experience.

Dedication

This project is dedicated to my beloved parents, Gerd Mella and Olav Mella, whose love and guidance are my constant sources of strength and inspiration. Your unwavering belief in me and your endless support have shaped the person I am today. This achievement is not only mine but also a testament to the sacrifices you have made and the wisdom you have imparted. I am forever grateful to walk this path with your blessings lighting the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Background

The broader landscape of data generation using Generative Adversarial Networks (GANs) presents its own set of challenges and opportunities. In recent times, GANs have emerged as a potent tool for generating complex data types, including images, speech, and text. However, the generation and evaluation of structured, high-dimensional tabular data present unique challenges. GANs, comprising a generator and a discriminator in adversarial training, excel in creating new data akin to a given dataset. Yet, the structured nature of tabular data, often stored in rows and columns like Excel spreadsheets, poses difficulties, especially when dealing with private sector data subject to stringent privacy regulations. The generation of high-quality synthetic tabular data is challenging due to the diversity of data types in columns and the complexity of non-Gaussian numerical data distributions. Consequently, there is no universally accepted generative model for tabular data, as different GAN models offer advantages in specific domains.

The integration of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare has been transformative, yet it faces critical challenges in data privacy and security. The previous work on Synthetic Tabular Data Generation (STDG) using CTGAN and CopulaGAN which are subsets of Generative Adversarial Networks (GANs) laid a foundation for addressing data scarcity and partial-privacy in healthcare. Building upon this, our study recognizes a pivotal gap: the need for enhanced robust data security in the generation of synthetic healthcare data. While GANs have proven effective in creating realistic synthetic datasets, the security of the original data during this process has not been sufficiently addressed. This thesis aims to fill this gap by exploring the application of Auto-Encoders (AEs) and Variational Auto-Encoders (VAEs), subsets of GAN, focusing on securing original healthcare data while maintaining the utility of the synthetic output.

1.2 Problem Statement

The primary problem addressed in this research is the vulnerability of original healthcare data when used in traditional synthetic data generation techniques, including GANs. Risks such as data leakage, unauthorized access, and potential reconstruction of original data pose significant threats to patient privacy and data integrity. This study seeks to explore the integration of AEs and VAEs in healthcare data generation, focusing on enhancing data security while maintaining the utility of the data and ensuring the confidentiality and integrity of the original datasets.

- How can AEs and VAEs be effectively integrated to enhance the security of original healthcare data in the synthetic data generation process?
- What impact does this integration have on the quality and utility of the generated

synthetic data?

- How can the balance between data security and clinical utility be optimized in synthetic healthcare data?

1.3 Research Objectives and Goals

The overarching objective of this research is to develop a secure and efficient framework for synthetic healthcare data generation that leverages the unique capabilities of AEs and VAEs. The specific goals include:

- To design and implement an integrated model that employs AEs and VAEs to enhance data security through sensitive data preprocessing.
- To evaluate the effectiveness of this model across various healthcare datasets, ensuring fidelity of data while generating high-quality synthetic data.
- To contribute to the advancement of secure and privacy-preserving synthetic data generation (SDG) in healthcare, focusing on the balance between data security, privacy, and clinical applicability.

1.4 Significance of the Study

This study is significant as it addresses the pressing need for secure artificial intelligence applications in healthcare. By enhancing data security through the innovative use of AEs and VAEs in synthetic data generation, this research contributes to the safe and ethical use of AI in healthcare, aligning with data protection laws and ethical standards. The findings aim to bolster trust in and applicability of synthetic healthcare data, facilitating its broader use in research and clinical practice.

For those interested in the detailed implementations, code, and datasets used in this research, all resources are comprehensively documented and made publicly available in the project's GitHub repository. The repository can be accessed at GitHub Repository for Enhancing Data Security with AE and VAE.

1.5 Scope of the Study

The scope of this thesis encompasses:

- Theoretical exploration of AEs and VAEs and their application in enhancing the security of synthetic data generation.
- Practical implementation and testing of the integrated model using diverse healthcare datasets.
- Empirical analysis of the model's performance in real-world scenarios, assessing data security, fidelity, and clinical utility.

1.6 Structure of the Thesis

This thesis is organized as follows: Chapter 2: Literature Review - A comprehensive review of relevant literature on AEs, VAEs, and data security in healthcare. Chapter 3: Methodology and Experimental Setup - Detailed description of the research methodology, including model design, and implementation strategies. And explanation of the experimental design. Chapter

4: Results and Data Analysis - Presentation and analytical techniques. Chapter 5: Discussion and Future Directions- Discussion of the research findings, their implications, and limitations. And suggestions for future research in this area. Chapter 6: Conclusion - Concluding remarks and

Chapter 2

Background

This chapter embarks on an extensive exploration of the foundational theories and related works pivotal to understanding and advancing the project's goals. The chapter begins by elucidating the core principles of neural networks, laying the groundwork for comprehending the intricate mechanisms of Generative Adversarial Networks (GANs). It further dissects the architecture of GANs, spotlighting the innovative variants that have been tailored for specific applications, including those in healthcare data synthesis. A significant focus is placed on the integration of Auto-Encoders (AEs) and Variational Auto-Encoders (VAEs) and integrating them into Conditional Tabular Generative Adversarial Networks (CTGAN) within the GAN framework, highlighting their critical role in enhancing data security and privacy. This discussion extends to the examination of various tabular GAN models, emphasizing their adaptability and effectiveness in generating synthetic healthcare data. The chapter also discusses the evaluation metrics and methodologies employed to assess the fidelity and utility of the synthesized data, ensuring it meets the stringent requirements of healthcare applications. Through a meticulous review of existing literature and emerging studies, this chapter aims to weave a comprehensive narrative that not only contextualizes the project within the broader field of AI in healthcare but also sets the stage for the innovative contributions this research aspires to make.

2.1 Neural Network Models

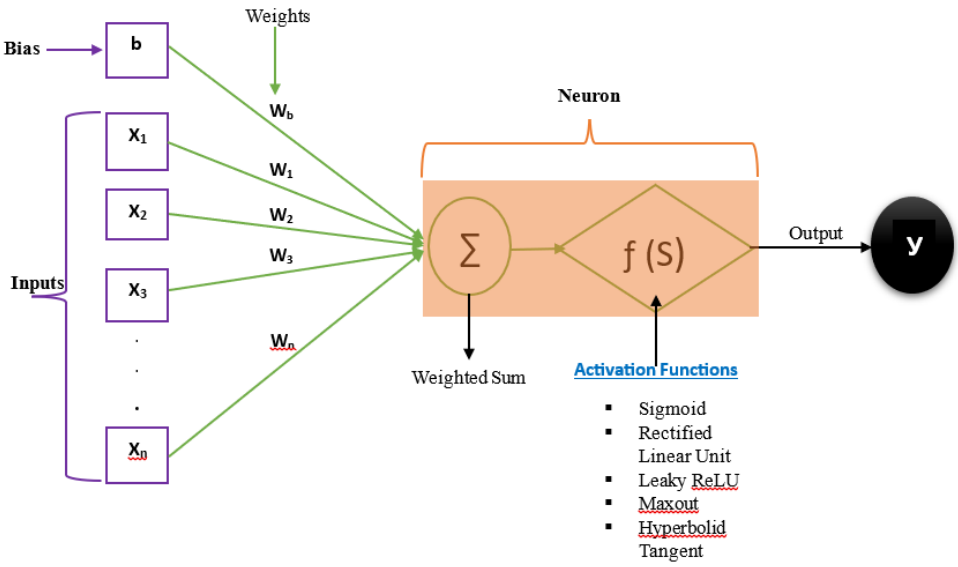


Figure 2.1: Basic neuron model structural framework.

Neural networks, a cornerstone of Artificial Intelligence (AI), emulate the brain's method

of processing information through a complex system of neurons, or nodes, that interact by sending electrical signals. These biological neurons generate outputs from inputs they receive, with the significance of each input determined by its weight [44, 68]. A basic neural network structure, depicted in Figure 2.1, demonstrates the core functionality of a neuron, laying the groundwork for the development of neural networks.

Basic Structure of an ANN.

At its core, an ANN is composed of nodes, or "neurons," arranged in layers: an input layer, one or more hidden layers, and an output layer. Each neuron in one layer connects to neurons in the next layer through pathways that are associated with weights, which are adjusted during the learning process.

- Input Layer: This is the entry point of the network. Each neuron in the input layer represents a feature of the input data. For example, in image recognition, each input neuron could represent a pixel's intensity.
- Hidden Layers: These layers perform the bulk of the computation through a network of neurons that process inputs from the previous layer. The hidden layers can extract and amplify features relevant to the task at hand, such as edges in visual data or semantic patterns in text.
- Output Layer: The final layer produces the network's output, such as a class label in a classification task or a value in a regression task. The structure of the output layer depends on the specific problem being addressed [44, 68].

Neuron Functionality.

Each neuron receives input from its predecessors, which it aggregates and transforms using an activation function. The activation function is crucial as it introduces non-linearity into the network, enabling it to learn complex patterns:

- Weighted Sum: A neuron calculates the weighted sum of its inputs, adding a bias term to account for input-independent adjustments.
- Activation Function: Common activation functions include the sigmoid, tanh, and ReLU (Rectified Linear Unit). These functions determine whether a neuron activates and to what extent, influencing the signal passed to subsequent neurons.

This basic neuron model, depicted in Figure 2.1, encapsulates a neuron's essential operations and lays the groundwork for constructing more complex neural network architectures.

2.1.1 Feed-Forward Neural Network

Feed-forward neural networks represent a fundamental architecture in neural network design, characterized by their ability to operate in both single-layer and multi-layer configurations. In its simplest form, a single-layer feed-forward network consists of just an input layer and an output layer. The input layer is tasked with receiving data, and the output layer is responsible for delivering the final computational results. Due to the network's linear, acyclic nature, information flows in a one-way direction from the input to the output neurons, without any feedback loops [78]. This straightforward structure, illustrated in Figure 2.2, underpins the basic operational framework of feed-forward networks.

2.1.2 Multi-layer feed-forward network model

A multi-layer feed-forward network enhances the basic architecture of a single-layer network by incorporating one or more hidden layers situated between the input and output layers. These hidden layers play a crucial role in performing complex computations and transformations on the data before it reaches the output layer. Through these intermediate layers, the network is capable of automatically generating features and applying necessary

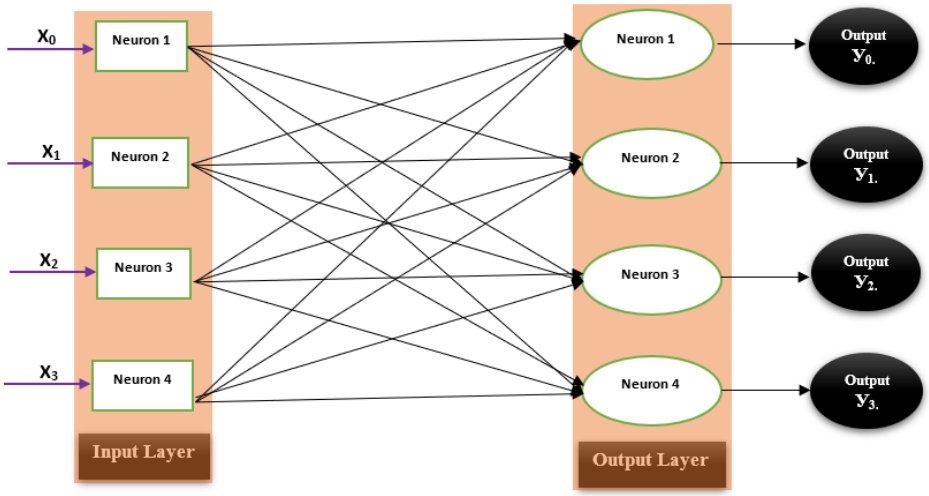


Figure 2.2: Single-layer feed-forward network.

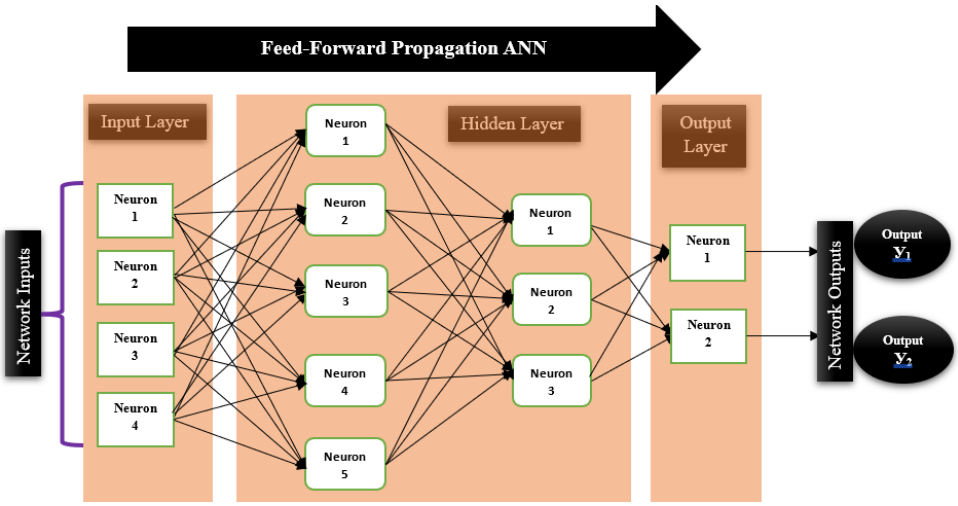


Figure 2.3: Multi-layer feed-forward network model.

transformations, thereby enriching the data processing capabilities of the model. The flow of information remains unidirectional, progressing sequentially from the input layer, through each hidden layer, and finally to the output layer. The complexity of the task at hand dictates the requisite number of hidden layers; simple tasks might be well-served by a single hidden layer, whereas more intricate problems could necessitate multiple hidden layers to achieve optimal performance [78]. The selection of the number of hidden layers and neurons is pivotal, as an incorrect configuration could lead to underfitting or overfitting, compromising the model's effectiveness. The intricacies of these issues and their implications will be further explored in Section 2.1.4.

2.1.3 Multi-layer back-propagation network model

For the backward propagation approach, also known as back-propagation, this method is integral to training multi-layer feed-forward networks. Back-propagation is a systematic way of updating the weights of the neurons in all layers, from the output back to the input layer, based on the error rate obtained in the output compared to the expected result. This process involves calculating the gradient of the loss function with respect to each weight by the chain rule, effectively measuring the impact of each weight on the loss. By iteratively adjusting the weights in the direction that minimizes the error, the network learns to make more accurate predictions. The back-propagation algorithm is crucial for optimizing the network's

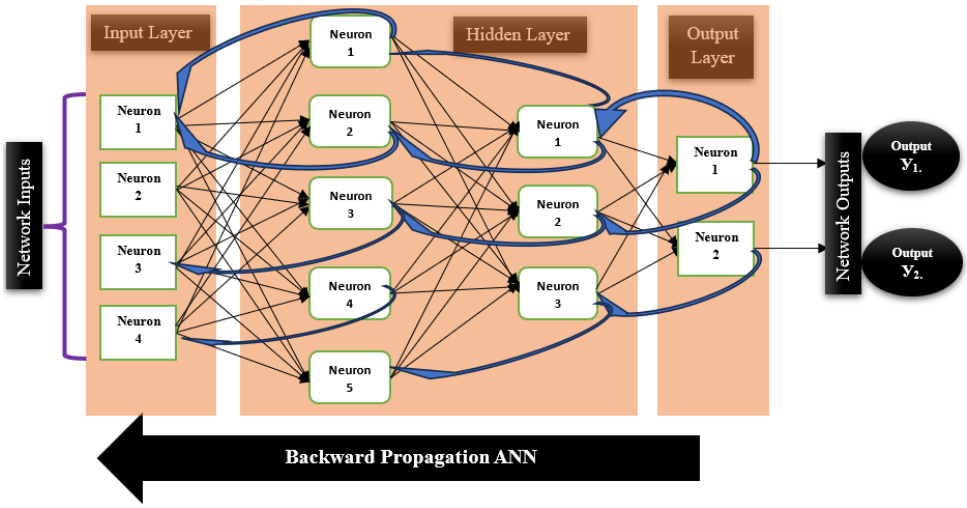


Figure 2.4: Multi-layer back-propagation network model.

performance, allowing it to learn complex patterns and relationships within the data. The depth and configuration of the hidden layers, which are meticulously adjusted during this process, are critical for the network's ability to solve complex problems without falling into the pitfalls of underfitting or overfitting [78].

2.1.4 Training Neural Networks

Training deep neural networks, characterized by their multiple hidden layers, relies on the backpropagation algorithm to refine the learning mechanism. This algorithm systematically tunes the network's weights and biases to reduce discrepancies between the predicted and actual outcomes. The essence of backpropagation lies in a two-phase cycle: a forward pass and a backward pass. Initially, input data propagate forward through the network, leading to the generation of predictions. These predictions are then evaluated against true values, with discrepancies quantified via a loss function. Subsequently, the computed error is disseminated backward across the network, prompting adjustments in weights and biases through an optimization algorithm. This backward pass leverages the gradient of the loss function, calculated with respect to each parameter via the chain rule, to guide the adjustments that aim to diminish the loss. This cyclical adjustment process persists until the network's predictions align closely with the actual data, signifying a reduction in error to a tolerable extent [68, 69]. The choice of an optimization algorithm plays a pivotal role in navigating the parameter space

towards the set of weights and biases that optimally reduce the error. The effectiveness of the training process hinges on this selection, with numerous algorithms at disposal. Among these, Stochastic Gradient Descent (SGD) and Adam stand out for their prevalent application in contemporary machine learning endeavors. These algorithms differ in their approach to adjusting parameters, with SGD focusing on updating parameters in a manner that minimizes the loss on a subset of the data at each iteration, and Adam enhancing this process by incorporating mechanisms that adjust the learning rate dynamically, accounting for the first and second moments of the gradients[25, 76].

Optimizing Neural Networks with Gradient Descent Techniques

In the realm of neural network optimization, gradient descent algorithms play a pivotal role by striving to minimize the cost function, guiding the model towards a local minimum. This optimization technique is versatile, encompassing batch, stochastic, and mini-batch variations, each tailored to different training dynamics.

- Batch Gradient Descent (BGD), also known as Vanilla Gradient Descent, operates on the principle of adjusting parameters by calculating the gradient of the loss function across the entire dataset. The update formula is given by:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{2.1}$$

where η represents the learning rate, a critical factor influencing the optimization speed and stability. A higher η accelerates convergence but risks bypassing the optimal solution, whereas a lower η ensures more accurate convergence at the expense of increased computation time. The gradient $\nabla_\theta J(\theta)$ directs the parameter adjustments, aiming to reduce the loss function $J(\theta)$. The comprehensive dataset usage in BGD, while thorough, renders it computationally intensive for large data volumes [70].

- Stochastic Gradient Descent (SGD) introduces a more dynamic approach by updating parameters for each individual training instance, thereby enhancing computational efficiency. The update mechanism is described as [70]:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^i, y^i) \tag{2.2}$$

In this context, x^i and y^i denote individual training inputs and their corresponding labels, making SGD faster but more susceptible to fluctuations, potentially affecting convergence precision.

- Mini-Batch Gradient Descent emerges as a balanced strategy, updating parameters in small subsets of the training data, defined by:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{i:i+n} : y^{i:i+n}) \tag{2.3}$$

This method combines the best of both worlds, leveraging batch efficiency and the stochastic approach's responsiveness, thus facilitating a smoother and more reliable path toward optimization [70]. Each variant of gradient descent offers unique advantages, with the choice largely dependent on the specific requirements of the neural network task at hand. The mini-batch approach is frequently favored for its equilibrium between efficiency and convergence stability, marking it as a versatile choice in neural network training endeavors.

Adam: An Advanced Optimization Technique for Neural Networks.

Adam, an acronym for Adaptive Moment Estimation, stands out as a sophisticated optimization algorithm that synergizes the strengths of Stochastic Gradient Descent (SGD) with momentum. It ingeniously incorporates the benefits of two prominent optimization methods, AdaGrad and RMSProp, to enhance neural network training efficiency while maintaining minimal memory usage [25, 76]. The hallmark of Adam is its ability to adjust learning rates for individual parameters dynamically through the computation of the gradients' first and second moments. This adaptability ensures that each parameter's learning rate is fine-tuned based on the historical gradients and their variability, a feature not present in traditional SGD, which applies a uniform learning rate across all updates [41]. Adam's methodological approach to adjusting learning rates on the fly makes it exceptionally adept at navigating the complex landscapes of neural network parameters. This adaptability not only accelerates the convergence process but also improves the overall training performance, especially in scenarios with intricate parameter spaces. Consequently, Adam is celebrated for its contribution to the efficient and effective optimization of neural networks, marking a significant advancement over conventional methods [41].

RMSprop: Enhancing Gradient Descent for Non-Stationary Objectives

RMSprop, short for Root Mean Square Propagation, is an optimization algorithm designed to resolve the diminishing or exploding learning rates issues encountered in AdaGrad. RMSprop modifies the AdaGrad approach to provide an adaptive learning rate that can be more suitable for dealing with non-stationary objectives, as often seen in recurrent neural networks and other deep learning models.
The update rule for RMSprop is given by:

1. Calculate the squared gradient:

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta) g_t^2 \tag{2.4}$$

2. Update the parameter:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{2.5}$$

Where:

- $E[g^2]_t$ is the moving average of the squared gradients at time step t,
- β is the decay rate, typically set to 0.9,
- g_t is the gradient at time step t,
- η is the learning rate,
- ϵ is a small scalar (e.g., $1e^{-8}$) to prevent division by zero.

The core idea behind RMSprop is to maintain a moving average of the squared gradients for each weight and to divide the learning rate by this average. This ensures that the learning rate is adjusted dynamically, allowing for larger updates for infrequent parameters and smaller updates for frequent ones. RMSprop's adaptive learning rate makes it particularly effective for problems where the optimal parameter space is complex and the gradients may change significantly across different training epochs[25, 41, 76].

Momentum: Accelerating SGD in the Relevant Direction

Momentum is a technique applied to the stochastic gradient descent (SGD) algorithm to accelerate the convergence towards the global minimum of the loss function. Incorporating a fraction of the update vector of the past steps to the current step helps in smoothing out the variations during training. This is akin to adding inertia to overcome the local minima and to speed up the descent in steady downhill directions[19, 52].

1. Update the velocity:

$$v_t = \gamma v_{t-1} + \eta g_t \tag{2.6}$$

2. Update the parameter:

$$\theta_{t+1} = \theta_t - v_t \tag{2.7}$$

Where:

- v_t is the velocity at time step t,
- γ is the momentum coefficient, close to 1 (e.g., 0.9),
- η is the learning rate,
- g_t is the gradient at time step t.

The momentum method accumulates an exponentially decaying moving average of past gradients and continues to move in their direction, increasing the speed of convergence. This approach not only helps in faster convergence but also aids in navigating through the rough terrain of the loss function landscape, making it a preferred choice in deep learning optimizations where the surface can be highly non-convex [19]. Both RMSprop and Momentum are pivotal in modern machine learning, offering unique advantages in optimizing neural networks. RMSprop addresses the issue of adaptive learning rate effectively, while Momentum accelerates the convergence by leveraging the direction of the previous gradients. Together with Adam and Gradient Descent, these algorithms form the backbone of optimization strategies in the field, each with its specific use cases and advantages[19, 41].

2.1.5 Activation Functions

In the realm of neural networks, activation functions are pivotal, serving as the gatekeepers of non-linearity, which is essential for the networks to capture complex patterns beyond what linear models can achieve [23]. Let's explore the essence and mathematical underpinnings of some fundamental activation functions, which are instrumental in enabling neural networks to understand and model intricate data relationships.

Sigmoid Function

The Sigmoid function, denoted as $\sigma(x)$, is a foundational activation function that maps any input value x to a range between 0 and 1, following the equation:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.8}$$

This characteristic makes it particularly suitable for binary classification tasks, especially in the output layer of simpler or shallow neural networks [23]. Despite its widespread use, the Sigmoid function can lead to challenges such as the vanishing gradient problem, particularly in deeper networks, as gradients of large magnitudes can become insubstantial during backpropagation, slowing down the learning process or halting it altogether.

Hyperbolic Tangent (Tanh) Function

The Tanh function, a scaled version of the Sigmoid, outputs values in a range of -1 to 1. It is mathematically expressed as:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.9}$$

Alternatively, it can be related to the Sigmoid function as:

$$\tanh(x) = 2\sigma(2x) - 1 \tag{2.10}$$

The Tanh function's output range allows it to handle data that spans both positive and negative values, making it more effective than the Sigmoid function in certain contexts, especially for hidden layers in a network [23].

Rectified Linear Unit (ReLU) Function

The ReLU function has become the default activation function for many types of neural networks due to its simplicity and efficiency [23]. It is defined as:

$$\text{ReLU}(x) = \max(0, x) \tag{2.11}$$

The derivative of the function, $g'(x)$, is defined as:

$$g'(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{2.12}$$

This piecewise function outputs x if x is positive and 0 otherwise. Its linear, non-saturating form facilitates faster convergence during training by mitigating the vanishing gradient problem [23]. However, ReLU can lead to **dead neurons** in which neurons stop responding to variations in error due to negative input values.

Leaky ReLU

To address the shortcomings of ReLU, the Leaky ReLU variant allows a small, non-zero gradient when the unit is not active, and x is less than zero:

$$\text{Leaky ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{if } x \leq 0 \end{cases} \tag{2.13}$$

Leaky ReLU ensures that all neurons remain active and updated during training, preventing the **dying ReLU** problem, and promoting more consistent learning across the network's architecture [23]. These activation functions are the linchpins in the design of neural networks, each with its own set of advantages and trade-offs, influencing the network's ability to learn and generalize from the data it's trained on.

2.1.6 Regularization Techniques

In the quest to optimize neural network performance, addressing the challenges of overfitting and underfitting is paramount. Overfitting is characterized by a model's exceptional performance on training data coupled with poor generalization to new, unseen data. This discrepancy often manifests as low training errors but elevated validation errors, indicating the model's tendency to memorize rather than learn from the training data. Conversely, underfitting occurs when a model fails to capture the underlying structure of the data, reflected in subpar performance on both training and validation datasets [43]. To mitigate these issues,

regularization techniques are employed, introducing modifications to the training process to enhance the model's generalization capabilities. Among the most prevalent regularization strategies are L1 and L2 regularization, dropout, and batch normalization, each with its unique approach to improving model robustness.

L1 Regularization (LASSO)

L1 regularization, or Least Absolute Shrinkage and Selection Operator (LASSO), imposes a penalty on the absolute magnitude of the model weights. This method encourages a sparse model with fewer weights, effectively reducing the model's complexity by driving non-critical feature weights to zero. The primary aim is to enhance feature selection, prioritizing the most influential features for the model's predictions [48].

L2 Regularization (Ridge/Tikhonov Regularization)

L2 regularization, also known as ridge or Tikhonov regularization, penalizes the square of the weights. Unlike L1, which can zero out weights entirely, L2 regularization tends to distribute the penalty across all weights, diminishing their magnitude without necessarily driving them to zero. This approach is less about feature elimination and more about ensuring no single feature dominates the model's output [63].

Dropout

Dropout is a dynamic regularization technique that randomly deactivates a subset of neurons during the training process. By temporarily removing neurons from the network, dropout prevents the model from becoming overly dependent on any specific set of features, promoting

more distributed and robust learning. This technique is particularly effective in reducing overfitting by simulating a wide variety of network architectures [63].

Batch Normalization

Batch normalization tackles the issue of internal covariance shift, where the distribution of inputs to layers shifts during training, complicating the learning process. By normalizing the inputs to each layer for every mini-batch, batch normalization stabilizes the learning environment, allowing for higher learning rates and more efficient training. This not only accelerates the training process but also contributes to improved model stability and performance on unseen data [38]. Regularization techniques are essential tools in the neural network toolkit, each contributing to the model's ability to learn from the data effectively without overfitting or underfitting. By carefully applying these techniques, practitioners can enhance the robustness and generalization of their neural network models, ensuring they perform well across a variety of datasets.

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) represent a paradigm shift in the field of machine learning, moving beyond the traditional confines of supervised learning, where models are trained on a predefined set of labeled data [18]. Supervised learning, despite its effectiveness in tasks such as classification and prediction, is inherently limited by its reliance on extensive, human-annotated datasets. This limitation has spurred interest in unsupervised learning approaches, which aim to reduce dependency on labeled data and human oversight.

Generative Modeling: A Leap into Unsupervised Learning Generative modeling, a subset of unsupervised learning, focuses on understanding and replicating the distribution of a given dataset. The objective is to model the underlying data-generating process, allowing for the creation of new data instances that are indistinguishable from real data. Generative Adversarial Networks (GANs) emerge as a powerful implementation of generative modeling, primarily known for their ability to synthesize highly realistic images [34].

2.2.1 Architecture and Training

Generative Adversarial Networks (GANs) embody a groundbreaking neural network framework that revolutionizes the way machines learn data patterns. This dual-network architecture, capable of functioning within both semi-supervised and unsupervised learning paradigms, showcases a dynamic interplay between two neural networks: the generator and the discriminator. The essence of GANs lies in their adversarial process, where the generator endeavors to fabricate data that mirrors the authenticity of real-world data, and the discriminator evaluates these creations alongside actual data to discern their genuineness [17]. The generator's goal is to produce images that closely resemble authentic images, while the discriminator evaluates both real and synthetic images to distinguish between the two. The generator improves its output based on feedback from the discriminator, which has access to both real and generated images. The discriminator's task is to accurately classify images as real or fake, using error analysis through backpropagation to guide the generator's improvements. The aim is to reach a point where the generator's images are indistinguishable from real images, achieving an equilibrium where the discriminator's accuracy is equivalent to random guessing.

During GAN training, both networks are optimized simultaneously. The discriminator aims to improve its classification accuracy, while the generator seeks to produce images that the discriminator will classify as real. This process is governed by the value function $V(G, D)$, defined as [33]:

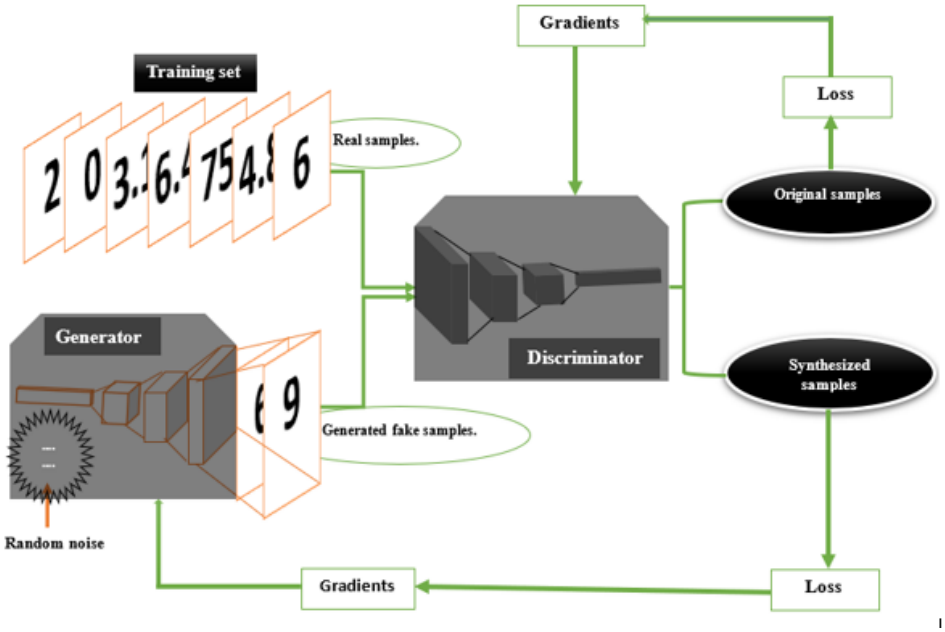


Figure 2.5: The architecture of Generative Adversarial Networks (GANs) showing the interaction between the generator and discriminator.

$$\max_D \min_G V(G,D) = \mathbb{E}_{p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{p_z(z)}[\log(1 - G(z))] \tag{2.14}$$

Here, G and D represent the generator and discriminator functions, respectively. p_z is the probability distribution of the latent space, and p_{data} is the probability distribution of the training dataset. The discriminator's goal is to maximize V by correctly identifying real and generated samples, while the generator aims to minimize V, indicating its success in deceiving the discriminator [33].
The discriminator is considered optimal when:

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \tag{2.15}$$

And the generator reaches optimality when the generated data distribution $p_g(x)$ matches the real data distribution $p_{\text{data}}(x)$, rendering the discriminator's accuracy no better than random guessing.

2.2.2 Navigating the Challenges of GAN Training

Training Generative Adversarial Networks (GANs) presents a unique set of challenges that can significantly impact their performance and stability. These challenges stem from the adversarial nature of GANs, where two networks, a generator and a discriminator, are trained simultaneously in a dynamic contest. This section investigates the common hurdles encountered during GAN training, including convergence issues, vanishing gradients, mode collapse, oscillatory loss, and the complexity of hyperparameter tuning [29, 51].

Convergence Difficulties: Achieving convergence in GANs is notoriously difficult due to the adversarial training process. Ideally, convergence signifies that the model has reached an optimal state where further training does not yield significant improvements. However, in the context of GANs, where the generator and discriminator are in a constant tug-of-war, maintaining a balance where both networks improve at a compatible rate is challenging. Non-convergence is often signaled by poor quality in the generated data, indicating that one network has outpaced the other, leading to stagnation in learning [62, 72].

Vanishing Gradients: A critical issue in GAN training is the vanishing gradient problem, particularly when the discriminator becomes too proficient early in training. This scenario results in the generator receiving minimal feedback to guide its improvements, effectively stalling its learning process. The discriminator's overwhelming accuracy creates a feedback void, preventing the generator from refining its output based on constructive criticism [29, 50].

Mode Collapse: Mode collapse occurs when the generator discovers a narrow path to deceive the discriminator, leading to a lack of diversity in the generated samples. This phenomenon restricts the generator to a subset of the data distribution, often producing repetitive or highly similar outputs. The root cause of mode collapse can be an overly dominant discriminator or a generator that settles for producing **safe** samples that are less likely to be classified as fake [15, 73, 77].

Oscillatory Loss: The training process of GANs can sometimes be characterized by oscillatory loss, where the loss metrics exhibit erratic fluctuations instead of stabilizing or showing consistent improvement. This instability complicates the training process, making it difficult to gauge the networks' progress and adjust training parameters effectively [29].

Hyperparameter Tuning: The complexity of GAN architectures necessitates meticulous hyperparameter tuning to achieve optimal performance. The vast array of hyperparameters, combined with the intricate interplay between the generator and discriminator, makes this task particularly daunting. Traditional methods like Grid Search are often impractical for GANs due to the computational demands and the nuanced impact of each parameter on the training dynamics [15, 29, 45, 62].

2.3 Exploring GAN Architectures

Beyond the foundational Vanilla GAN introduced by [80, 87], the field has seen the emergence of diverse GAN architectures, each tailored to overcome specific challenges or to enhance performance in generating complex data distributions. This section highlights significant advancements in GAN architectures, focusing on their unique features and applications.

2.3.1 Conditional GAN (CGAN)

Introduced by Mirza and Osindero in 2014, Conditional GANs (CGANs) emerged to overcome the limitations of Vanilla GANs, particularly their inability to direct the generation process [56]. CGANs integrate additional information, such as class labels or other relevant data, into the training phase, enabling a more targeted generation of data. This approach allows CGANs to produce data specific to predetermined categories, enhancing the versatility of GANs for various applications. The modification involves adding a conditional variable y to both the generator and discriminator, leading to a new training equation:

$$V(G, D) = \mathbb{E}_{pdata(x)}[\log D(x|y)] + \mathbb{E}_{p_z(z)}[\log(1 - G(z|y))] \tag{2.16}$$

This advancement has paved the way for CGANs to be applied in tasks like image-to-image translation and more, where conditional input plays a crucial role in shaping the output.

2.3.2 Deep Convolutional GAN (DCGAN)

To address the challenge of generating high-quality images, Radford et al. introduced Deep Convolutional GANs (DCGANs) in 2015. By incorporating Convolutional Neural Networks (CNNs) into both the generator and discriminator, DCGANs significantly improve the quality of generated images. CNNs, known for their effectiveness in image-related tasks, analyze input images through layers of convolution and pooling to identify and extract features [75]. DCGANs enhance this process with architectural guidelines such as replacing pooling layers

with down-sampling convolutions in the discriminator, using deconvolutions in the generator, and applying batch normalization to both. These guidelines, along with the use of specific activation functions (ReLU in the generator and LeakyReLU in the discriminator), contribute to the improved stability and performance of DCGANs in image generation tasks [66].

2.3.3 Wasserstein GAN

Introduced in 2017 by Arjovsky and colleagues [2, 3], the Wasserstein Generative Adversarial Network (WGAN) marks a significant advancement in stabilizing GAN training and mitigating the issue of mode collapse, where the generator tends to produce a narrow range of outputs. Unlike traditional GANs, WGAN introduces a novel loss metric, the Earth Mover's Distance (EMD) or Wasserstein metric, enhancing the interpretability of learning curves and aiding in the fine-tuning of model parameters. This metric measures the discrepancy between the distribution of real and generated data without relying on binary classification, offering a more nuanced assessment of model performance.

In WGAN, the discriminator, referred to as the **critic**, evaluates the EMD, aiming to quantify the effort required to transform the generated data distribution into the real data distribution. The goal of the generator is to minimize this distance, effectively making the generated data indistinguishable from real data. This approach not only provides a clearer understanding of the model's loss but also simplifies hyperparameter optimization by incorporating a clipping constraint to regulate the training process, offering a straightforward criterion for determining convergence.

The introduction of the 1-Lipschitz constraint further stabilizes training by ensuring the critic's output remains within a bounded gradient, reducing the risk of vanishing gradients and overfitting [7, 46]. This technical adjustment allows the critic to offer more detailed feedback to the generator, promoting the production of varied outputs that more accurately reflect the diversity of the target distribution [28]. Consequently, WGAN addresses the challenge of mode collapse, facilitating the generation of a broader array of samples and enhancing the overall robustness of GAN training

2.3.4 Wasserstein GAN with Gradient Penalty (WGAN-GP)

The WGAN with Gradient Penalty (WGAN-GP) improves upon the original WGAN by introducing a gradient penalty to enhance training stability and sample quality, addressing issues like low-quality outputs and convergence difficulties. This adjustment avoids explicit weight clipping through a Lipschitz constraint, with a gradient penalty coefficient λ recommended at $\lambda = 10$, subject to model and dataset specifics. WGAN-GP diverges from common GAN practices by employing layer normalization in the critic, leading to better stability and performance without extensive hyperparameter tuning.

2.4 Generating Synthetic Tabular Data Using GANs

For Synthetic Tabular Data Generation (STDG), GANs offer a robust alternative to traditional statistical models, especially for large datasets. They are trained to mimic the data distribution across a dataset's columns, producing synthetic tables that closely resemble real data and can handle diverse data types including discrete, continuous, and categorical. Despite being an emerging field, GANs for STDG have shown promise, with various approaches being explored to overcome challenges and effectively evaluate synthetic data quality[62].

2.4.1 Highlight of Tabular GAN Methods

Reviews by [36] and [16] highlight the superiority of GAN-based methods in Synthetic Tabular Data Generation (STDG) but do not single out a definitive best approach. The effectiveness of GAN models varies with dataset specifics, necessitating experimentation to identify the most suitable architecture. Despite challenges in balancing resemblance, privacy, and utility, several notable GAN models have emerged:

- **MedGAN** [14] in 2017, pioneer in synthesizing medical data, leveraging an autoencoder within its GAN framework to adeptly handle binary and count data types, aiming for a realistic emulation of patient records.
- **MedWGAN and MedBGAN** [5] in 2018, enhance MedGAN's capabilities by integrating WGAN-GP for MedWGAN and a boundary-seeking GAN for MedBGAN, respectively, elevating the realism in patient record generation.
- **MedGAN** [9] 2018 innovation refines MedGAN to generate superior multicategorical data, incorporating Gumbel softmax layers, thus broadening the scope of data types the model can accurately replicate.
- **HealthGAN** [84], introduced by Yale et al. in 2020, overcomes MedGAN's constraints, particularly in handling categorical data, by incorporating WGAN-GP, thereby improving the model's versatility and accuracy in data synthesis.
- **TableGAN and TGAN**, developed by [59] and [82] respectively, are distinguished by their focus on producing synthetic records across a spectrum of data types, with TableGAN utilizing CNNs for structure and TGAN employing RNNs to enhance sequential data generation.
- **CTGAN and CopulaGAN** [60, 62, 83], represent evolutionary steps forward, introducing sophisticated techniques for condition-specific data generation, trained with WGAN-GP, thus enabling more targeted and nuanced synthetic data creation.
- **TimeGAN** [85], specifically tackles the intricacies of time-series data, integrating autoencoding mechanisms to preserve temporal dynamics, offering a nuanced approach to time-series synthesis.
- **C-TABGAN and TabFairGAN** [67, 88], the latest in the series, not only push the envelope in data type diversity and handling imbalances but also embed fairness directly into the GAN architecture, setting new standards for ethical synthetic data generation.

These models demonstrate the evolving landscape of GAN-based STDG, each contributing unique solutions to the challenges of generating realistic, privacy-conscious synthetic tabular data.

2.4.2 Issues in Generating Synthetic Tabular Data with GANs

Tabular GANs face unique challenges when generating synthetic data, particularly from complex datasets like Electronic Health Records (EHRs), which often exhibit class imbalance [30]. This imbalance can skew the model's performance, leading to biased predictions. Additionally, the diversity in data distribution across columns can cause issues such as non-convergence and vanishing gradients [8]. Sparse data and one-hot encoding further complicate training, as the discriminator may struggle to distinguish between real and synthetic data based on rarity rather than authenticity [8].

To mitigate these issues, it's crucial to utilize diverse and high-quality datasets, including those directly sourced from healthcare providers, to ensure a broad representation of real-world complexities. However, the lack of standardized documentation and potential quality issues in open-source datasets pose significant challenges to data reliability and model validity [31].

Tabular GAN Approach	Year	Underlying Model Framework
MedGAN [14, 62]	2017	Vanilla GAN, Auto-encoder
MedWGAN [5, 62]	2018	WGAN-GP
MedBGAN [5, 62]	2018	BGAN
MedGAN with Gumbel-softmax [9, 62]	2018	Traditional GAN, Auto-encoder
HealthGAN [62, 84]	2020	WGAN-GP
TableGAN [59, 62]	2018	DCGAN
TGAN [62, 82]	2018	Traditional GAN
CTGAN [62, 83]	2019	WGAN-GP
CopulaGAN [60, 62]	2019	WGAN-GP
TimeGAN [62, 85]	2019	RCGAN, C-RNN-GAN
C-TABGAN [62, 88]	2021	CGAN
TabFairGAN [62, 67]	2022	WGAN

Table 2.1: Highlight of common Tabular GAN methods and the key framework composition.

2.4.3 Exploring CTGANs for Synthetic Tabular Data Generation

The invention of Conditional Generative Adversarial Networks (CTGAN) can be attributed to the researchers Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni from the Massachusetts Institute of Technology (MIT). Their work, titled "Modeling Tabular Data using Conditional GAN," was presented and elaborated upon in their research paper. This work laid the foundation for CTGAN, introducing a novel approach to generating synthetic tabular data, particularly effective in handling the complexities associated with categorical and imbalanced data[83]. It can be conditioned on specific data features, enabling the generation of synthetic data that mirrors the conditional distributions of the original dataset.

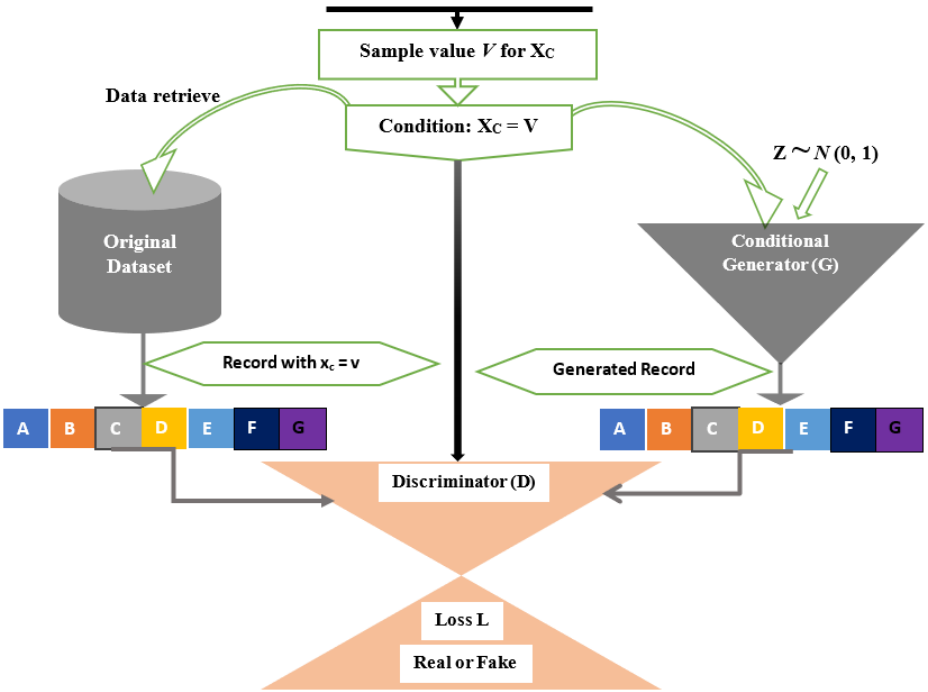


Figure 2.6: The Conditional Tabular Generative Adversarial Networks (GANs) showing the interaction between the generator and discriminator.

To complement the visual depiction of a Conditional Tabular Generative Adversarial Network (CTGAN), here's a breakdown of the CTGAN components as depicted in the above figure:

1. **Sample Value V for Xc**: This represents a specific condition or value for a categorical

column Xc in the dataset. For instance, if Xc is the "Gender" column, V could be "Male". This condition guides the generation process to produce data that adheres to this specific characteristic.

2. **Condition: Xc = V**: This condition serves as an input to both the generator and discriminator, indicating that the synthetic data generation should focus on records matching Xc = V. It ensures that the generated data respects the specified condition, allowing for targeted data augmentation.

3. **Arrow for Data Retrieving**: This arrow points from the condition Xc = V to the original dataset, signifying the process of selecting or considering real data instances that match the given condition. This step is crucial for learning the distribution of data conditioned on Xc = V.

4. **Conditional Generator (G) with Z∼N (0,1)**: The generator G takes noise Z drawn from a normal distribution N (0,1) and the condition Xc = V as inputs. It then generates synthetic records that not only resemble the real data but also satisfy the specified condition Xc = V.

5. **Arrow Pointing to Discriminator (D) with a Generated Record**: This arrow indicates the flow of synthetic data from the generator *G* to the discriminator *D*. The discriminator's role is to distinguish between real data instances from the original dataset and fake instances generated by *G*.

6. **Loss L 'Real or Fake'**: This phrase under the discriminator *D* represents the outcome of the discrimination process. The loss L quantifies how well *D* can differentiate real data from fake. During training, *G* aims to minimize this loss by improving its ability to generate data that *D* cannot easily classify as fake.

7. **Arrow from Original Data to Discriminator (D) with Record Xc = v**: This represents the flow of real data instances that satisfy the condition **Xc = V** into the discriminator *D*. It allows *D* to learn the characteristics of real data that matches the specified condition.

8. **Arrow from Condition: Xc = V to Discriminator (D)**: This indicates that the discriminator also receives the condition **Xc = V** as part of its input, allowing it to evaluate whether the records (both real and synthetic) align with the specified condition.

CTGAN leverages the adversarial relationship between the generator G and the discriminator D to produce high-quality synthetic tabular data that is conditioned on specific features. By incorporating conditions directly into the generation process, CTGAN ensures that the synthetic data not only mimics the overall distribution of the original dataset but also adheres to specific characteristics defined by the user. This capability makes CTGAN an invaluable tool for data augmentation, privacy preservation, and overcoming data imbalances in machine learning applications.

2.4.4 Dissimilarity and Similarity Between STDG and CTGAN

Synthetic Tabular Data Generation (STDG) and CTGAN (Conditional Tabular Generative Adversarial Network) both address the challenge of generating synthetic tabular data, but they do so in different manners. Here's a breakdown of their key differences and similarities, as well as insights into their usage and underlying algorithms:
Key Dissimilarity

- **Definition and Scope:**

    - STDG is a broad term that encompasses various methodologies and algorithms used to generate synthetic data resembling original tabular datasets. It includes a wide range of techniques, from simpler statistical methods to more complex machine learning models.
    - CTGAN is a specific type of GAN (Generative Adversarial Network) designed to generate synthetic tabular data. It addresses specific challenges of tabular data, such as handling mixed data types (continuous and discrete variables) and imbalanced data.

- **Algorithm Specificity:**
    - STDG does not refer to a specific algorithm but rather a category of methods, which can include anything from basic duplication with noise addition to sophisticated deep learning models.
    - CTGAN specifically utilizes a generative adversarial network architecture with modifications to better suit tabular data, such as using conditional vectors to handle category imbalance and mode-specific normalization for continuous columns.

Key Similarities

- **Objective:** Both STDG and CTGAN aim to generate new tabular data points that maintain the statistical properties of the original dataset. This includes preserving relationships between variables, distributions of individual columns, and potentially the privacy of the data subjects.
- **Use Cases:** They can be used for similar purposes, such as data augmentation (increasing the size of a dataset), privacy-preserving data sharing, imbalanced dataset handling, and synthetic dataset creation for testing machine learning models.

Can They Be Used the Same Way?
While STDG and CTGAN can be used for similar end goals, the approach and implementation details will differ based on the chosen STDG technique. CTGAN offers a more specific, ready-to-use solution for synthetic tabular data generation, especially beneficial for complex datasets with mixed types of data and class imbalances. STDG techniques might require more customization or selection of appropriate methods based on the dataset's characteristics and the specific requirements of the task at hand.

Do They Share the Same Algorithm?
No, CTGAN is a specific algorithm within the broader category of STDG methods. STDG encompasses a wide range of algorithms, including but not limited to GAN-based approaches like CTGAN. Other STDG methods might employ different machine learning models, statistical techniques, or data manipulation strategies that do not utilize the adversarial training approach central to CTGAN.

While both STDG and CTGAN serve the purpose of generating synthetic tabular data, CTGAN provides a targeted approach with innovations specifically designed to address the nuances of tabular datasets. STDG, being a broader term, offers a wider array of methods that can be tailored to various data generation needs, not all of which will leverage the advanced features or require the complexity of CTGAN. The choice between using a specific method like CTGAN or another STDG technique depends on the dataset's specific challenges, the desired level of sophistication in handling those challenges, and the expertise available to implement these solutions.

2.4.5 Evaluation Metrics for Synthetic Data Generation

Evaluating the effectiveness of Generative Adversarial Networks (GANs) in creating synthetic tabular data involves three key dimensions: resemblance to original data, utility in machine

learning tasks, and privacy protection. These metrics offer insights into the synthetic data's quality and its applicability in various domains.

Comparative Analysis

Comparative analysis measures resemblance or how closely synthetic data mirrors the statistical properties of the original dataset. This involves comparing basic statistical measures (mean, median, standard deviation) and employing more sophisticated tests to ensure the synthetic data accurately reflects the original data's distribution [35].

- **Statistical Comparisons**: Direct comparison of univariate statistics such as mean, median, and standard deviation between synthetic and real data.
- **Dimension-Wise (DW) Testing**: Evaluates each feature's distribution by comparing synthetic and real data using probability and distance metrics. Common tests include the Bernoulli's success probability, chi-squared tests for binary features, and the Student T-test for continuous features.
- **Distance Metrics**: Tools like Cosine Distance measure the similarity between the feature distributions of synthetic and real data, with smaller distances indicating better resemblance.
- **Joint Distribution Analysis**: Ensures that the synthetic data maintains the real data's inter-feature relationships, using metrics like Jensen-Shannon Divergence and Wasserstein Distance.
- **Correlation Analysis**: Assesses whether synthetic data preserves the real data's inter-dimensional relationships and correlations, employing Pearson and Spearman correlation coefficients among others.
- **Visualization**: Distribution plots and principal component analysis (PCA) visually compare the real and synthetic data, highlighting their statistical and structural similarities.

Data Visualization Techniques

Data Visualization Techniques refer to a broad range of methods and tools used to graphically represent complex data sets, making them easier to understand and interpret. These techniques transform numerical and categorical data into visual formats, such as charts, graphs, plots, and maps, allowing users to observe patterns, trends, anomalies, and correlations within the data. Effective data visualization aids in data analysis, decision-making, and communication of insights to both technical and non-technical audiences.

- **Histogram**: Histograms are graphical representations of data distribution, grouping data into bins to visualize frequency and distribution characteristics.
- **Density Plots**: Density Plots provide a smooth, continuous version of histograms, showing data distribution over an interval to identify concentration areas.
- **Box Plots**: Box Plots offer a visual summary of data's key quartiles, including the median, the 25th and 75th percentiles, and outliers, useful for comparing distributions and spotting outliers.
- **Scatter Plots**: A scatter plot uses dots to illustrate the relationship between two variables, with each dot representing a data point. It's effective for spotting trends, correlations, and outliers by observing the arrangement and direction of the dots.
- **Heat Maps**: A heat map uses color gradients to represent values, making it easy to visualize complex data patterns. It's useful for identifying correlations and trends, with color intensity indicating the magnitude of values.

Sensitivity Analysis and Utility Measurement

The utility of synthetic data in machine learning (ML) tasks is assessed by its ability to substitute for real data without significant loss in model performance. This dimension evaluates whether models trained on synthetic data can achieve comparable results to those trained on real data [29, 36].

- **Train on Synthetic, Test on Real (TSTR)**: A framework where ML models are trained on synthetic data and tested on real data. High performance indicates the synthetic data's effectiveness in capturing the original data's underlying distribution [27, 31].
- **Train on Real, Test on Synthetic (TRTS)**: This involves training a model on real data and testing it on synthetic data, assessing the utility of synthetic data for model validation[31].
- **Train on Synthetic, Test on Synthetic (TSTS)**: TSTS, where both training and testing are done on synthetic data, evaluates the internal consistency and reliability of synthetic data for machine learning.
- **Synthetic Ranking Agreement (SRA)**: This method evaluates how well synthetic data maintains the ranking order of predictive models based on their performance on real data [40].
- **Data Combination**: Examines the impact of synthetic data when used to augment real datasets, enhancing the diversity and size of training data for ML models [79].
- **Performance Metrics**: Common metrics such as Area Under the Curve (AUC), F1-score, and Accuracy are used to quantify ML models' performance on synthetic versus real data, providing a quantitative measure of the synthetic data's utility [16].

Privacy Risks Assessment

Privacy risk assessment is a process to identify and evaluate the risks of re-identification in shared or published datasets. It's vital to ensure that individual's privacy is maintained while allowing data utility, such as medical records.

- **T-Closeness**: This technique addresses a limitation of k-anonymity and l-diversity by ensuring that the distribution of a sensitive attribute within any group is no more than a threshold t away from the distribution of the attribute in the entire dataset. This helps protect against attacks that exploit the distribution of sensitive attributes.
- **K-Anonymity**: K-Anonymity ensures that each record in a dataset is indistinguishable from at least $k-1$ other records with respect to certain "quasi-identifier" attributes. This provides a measure of anonymity by preventing individual data from being singled out.
- **L-Diversity**: L-Diversity extends k-anonymity, requiring that the distribution of sensitive attributes in each group has at least l "well-represented" values. It protects against attacks leveraging insufficient diversity within anonymized groups.
- **Differential Privacy (DP)**: The gold standard for privacy protection in synthetic data, DP ensures that the removal or addition of a single data point does not significantly alter the outcome of data analysis, thereby protecting individual privacy [26].
- **Attack Methods**: Evaluating privacy also involves testing the synthetic data against potential attack methods, including membership inference attacks, attribute disclosure attacks, and model inversion attacks. These methods assess the robustness of synthetic data against attempts to extract or infer sensitive information [11, 58, 86].

- **Additional Privacy Measures**: Other techniques, such as Euclidean distance checks, exact match searches, and K-Nearest Neighbours (KNN) analysis, provide further insights into the privacy aspects of synthetic datasets [16, 31].

Evaluation Criteria Overview

Figure 2.7 below summarizes the evaluation metrics and methods used in the literature for assessing synthetic datasets across the four key criteria as listed in the figure:

Assessment Criteria	Evaluation Type	Techniques Used
Comparative Analysis	Dimension-wise testing	Bernoulli X^2 test Student T-test Mann-Whitney U-Test KS Test F-Test Cumulative distribution Cosine Distance
	Joint-distribution similarity	JS-Divergence Inception Score Wasserstein Distance Maximum Mean Discrepancies
	Inter-dimensional relationship similarity	Pearson Correlation, Spearman Correlation, Correlation coefficients, Correlation matrices, Dimension-Wise prediction tests
	Graphical Comparison	Histogram Density Plots Box Plots
Sensitivity Analysis	Model Dependence	Rerun Models with Original and Synthetic Data with TSTR, TRTS, TSTS, SRA
Utility Measurement	Classification	F1-score Recall Precision Accuracy
	Regression	AUC AUPRC MRE
Privacy Risk Assessment	Re-Identification Risk	K-Anonymity I-Diversity T-Closeness
	Attacks	Membership Inference Attribute disclosure. Model Inversion
	Other techniques	Euclidean Distance KNN Exact Matches

Figure 2.7: Assessment Criteria and Evaluation Techniques

Clinical Evaluation and Performance Considerations

- **Clinical Evaluation:** The involvement of health professionals in assessing synthetic datasets is crucial but often overlooked. Only a minority of studies engage clinicians to validate the practicality and reliability of synthetic patient data [16].

- **Performance and Computational Cost:** The efficiency of synthetic data generation processes, in terms of computational resources and time, is rarely analysed. Yet, it's essential for practical applications, especially in resource-constrained settings [35, 36].
- **Privacy vs. Data Similarity:** Achieving a balance between maintaining privacy and ensuring high data fidelity is challenging. Privacy measures often reduce the resemblance to the original data, posing a dilemma for dataset creators [31].
- **Standardization of Metrics:** There's a lack of standardized benchmarks for evaluating synthetic datasets, particularly in the medical field. This gap leads to inconsistencies and challenges in comparing different synthetic data generation methods [31].

In conclusion, while generating high-quality synthetic datasets is feasible, incorporating robust privacy protections without compromising data utility remains a complex challenge. A multidimensional evaluation approach, involving both technical metrics and clinical validation, is essential for advancing the field of synthetic data generation.
Each of these techniques and tools in the above discussions plays a specific role in data analysis, privacy protection, and the evaluation of synthetic data, ensuring valuable insights while respecting privacy.

2.5 Advancements and Hurdles in Synthetic Health Data Generation

The generation of synthetic health data stands at the forefront of healthcare research, tackling the critical issue of data scarcity amidst strict regulatory frameworks. This initiative is entangled with ethical concerns, notably the protection of patient privacy and the mitigation of biases, highlighting the complexities of applying generative models in healthcare. The process of data sampling requires rigorous scrutiny to avoid sample-selection biases, which could compromise the generalizability of models. Such biases emerge when models disproportionately represent certain demographics or are trained exclusively on data from specific types of medical equipment, leading to outcomes that fail to generalize across different patient populations [12]. Furthermore, class imbalance exacerbates biases, particularly in the diagnosis and prognosis of rare diseases, emphasizing that the effectiveness of AI models is contingent upon the quality and diversity of their training data [12]. The integration of AI algorithms in medical devices is increasing, yet the reliance on current generative models introduces concerns regarding patient privacy and data integrity. Vulnerabilities like membership inference attacks underscore the risks associated with synthetic data, where attackers could leverage model information to compromise patient confidentiality [12].

In 2021, Chen et al. [12] highlighted the challenges in adopting synthetic data within healthcare, proposing its use as a provisional solution for model enhancement. However, the lack of clinical benchmarks for assessing the quality of synthetic data, especially for rare diseases, presents significant challenges. The difficulty in interpreting evaluation metrics by clinicians calls for the development of assessment tools that are more understandable to healthcare professionals, to foster trust in the use of synthetic data [31]. The concept of visual Turing tests for evaluating synthetic image data, although thorough, proves impractical for large datasets and is less relevant for tabular health records due to their abstract nature [31]. Expanding data collection across various healthcare institutions could improve model robustness and generalizability. However, this approach faces challenges related to data sharing regulations, which often conflict with data protection laws [12].

This section primarily addresses the use of synthetic data to augment real data for model refinement and privacy concerns. Nonetheless, synthetic data harbours potential for broader applications, including stress-testing AI algorithms, simulating diverse scenarios in virtual environments, and training AI models on surgical errors without risking patient safety [12].

2.6 Autoencoders (AEs)

The concept of autoencoders was born by Rumelhart, Hinton, and Williams in 1986, aiming to minimize reconstruction errors, essentially learning to replicate the input data as closely as possible. Autoencoders have been applied in various domains, including dimensionality reduction, feature learning, and anomaly detection. They are a subset of neural network architectures used to learn efficient representations of input data, called encodings, in an unsupervised manner. The primary goal of an autoencoder is to compress the input data into a latent-space representation and then reconstruct the input data as accurately as possible from this representation. This process involves two main components: the encoder, which compresses the input, and the decoder, which reconstructs the input from the compressed form back to its original dimension [71].

Unlike traditional supervised learning, where each input x_i is paired with a corresponding label y_i, autoencoders deal with datasets S_T comprising solely of inputs:

$$S_T = \{x_i | I = 1, \ldots, M\} \tag{2.17}$$

Here, $x_i \in \mathbb{R}^n$ represents an observation in an n-dimensional space, and M is the total number of observations. Their ability to learn compressed representations makes them particularly useful for enhancing data security, as the latent representation can be manipulated to remove or alter sensitive information before reconstruction, thereby generating synthetic data that preserves privacy.

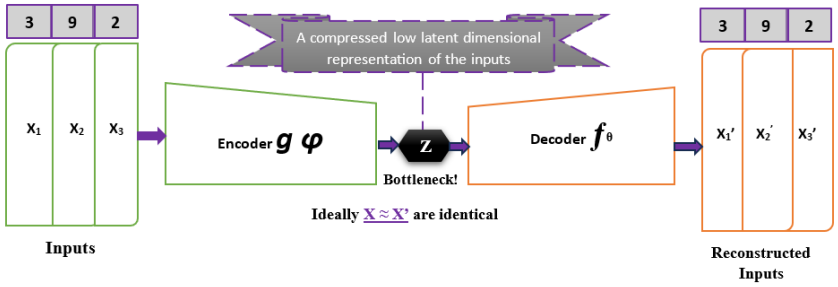


Figure 2.8: Simple Autoencoder Architecture

The architecture of an autoencoder is divided into three key components:

1. **Encoder:** This function maps the input data to a latent space, effectively compressing the data into a lower-dimensional representation[71].

2. **Latent Space (Feature Representation):** This is the core of the autoencoder, where the data is represented in a compressed form. This latent space holds the "encoded" information of the input data[71].

3. **Decoder:** The decoder function attempts to reconstruct the input data from the compressed latent representation[71].

The encoder and decoder are neural networks themselves, and the latent space is typically represented as a tensor of real numbers. The challenge and the art of designing an autoencoder lie in ensuring that the latent space representation is both "informative" and "useful" for further applications, such as feature extraction, data compression, or even generative tasks. Autoencoders, particularly when structured with neural networks, offer a sophisticated approach to learning data representations by encoding inputs into a compressed latent space and subsequently reconstructing them. This process, facilitated by libraries like TensorFlow or PyTorch, leverages backpropagation for efficient training[71]. The encoder function, denoted as g, transforms the input x_i into a latent representation h_i, where $h_i \in \mathbb{R}^q$, effectively reducing the data to its essential features within a lower-dimensional space:

$$h_i = g(x_i) \tag{2.18}$$

The decoder, represented by function f, then aims to reconstruct the input from this compressed form:

$$\tilde{x}_i = f(h_i) = f(g(x_i)) \tag{2.19}$$

The training objective of an autoencoder is to minimize the discrepancy between the original input x_i and its reconstruction $\tilde{x}_i$, typically using a loss function that penalizes differences between the two. This is formally expressed as:

$$\arg\min_{f,g} \langle L(x_i, f(g(x_i))) \rangle \tag{2.20}$$

where L signifies the loss function, measuring the reconstruction error, and the angle brackets denote averaging over all samples. To prevent the autoencoder from merely learning an identity function — which would be trivial and not particularly useful — two main strategies are employed: introducing a bottleneck and applying regularization.

Regularization

Regularization introduces additional constraints or penalties on the model, encouraging it to learn more robust and generalizable features. This can be achieved through various means, such as penalizing the magnitude of the weights (L1/L2 regularization), encouraging sparsity in the latent representations, or using dropout to prevent over-reliance on specific paths through the network. By carefully designing the encoder and decoder functions (g and f, respectively) and incorporating strategies like bottlenecks and regularization, autoencoders can learn to compress and reconstruct data effectively. This process not only aids in data compression but also in learning representations that are useful for tasks such as anomaly detection, denoising, or even generative modelling. The goal is to achieve a balance where the autoencoder reconstructs the inputs accurately while also discovering meaningful and useful patterns in the data [4, 55].

Bottleneck Approach

The most straightforward regularization technique involves creating a bottleneck by reducing the dimensionality of the latent space to be smaller than that of the input space. This forces the autoencoder to prioritize which aspects of the input data are most important to retain, leading to a more meaningful and compressed representation. Such representations are invaluable for tasks like feature extraction, data compression, and more. However, it's worth noting that even with a single bottleneck node, overfitting can occur if the encoder and decoder have enough capacity to map each input sample to a unique code in the latent space[4, 55].

Beyond Bottlenecks: Advanced Regularization Techniques

When the latent space is large, preventing the autoencoder from defaulting to an identity function requires more sophisticated forms of regularization. These methods aim to encourage the autoencoder to learn useful features about the data rather than memorizing it:

- **Sparse Autoencoders:** By adding a sparsity constraint on the activations of the hidden layers, sparse autoencoders encourage the model to represent each input with a small number of active neurons in the bottleneck layer. This can lead to more distinctive and interpretable representations[4, 55]. The key methods are:

    i. **L1 Regularization:** A framework where ML models are trained on synthetic data and tested on real data. High performance indicates the synthetic data's effectiveness in capturing the original data's underlying distribution.

ii. **KL-Divergence:** The method adjusts the sparsity level by manipulating the expected activation probability p of each neuron. The empirical probability p_j for each neuron j is calculated across a batch, and the KL-divergence between p and p_j is minimized as part of the loss function.

- **Denoising Autoencoders:** These autoencoders are trained to reconstruct the original input from a corrupted version, which forces the model to learn more robust features of the data. The corrupted input $\tilde{x}$ follows a distribution $C(\tilde{x}|x)$, with common choices being Gaussian noise $N(x, \sigma^2 I)$ or dropout noise Bernoulli(p), where p is the probability of an input element being kept[4, 55].

- **Contractive Autoencoders:** These autoencoders focus on making the learned representations robust to small input perturbations by penalizing the sensitivity of the hidden representations to changes in the input. This is achieved by minimizing the Frobenius norm of the Jacobian matrix of the encoder's activations with respect to the input, effectively encouraging the encoder to ignore irrelevant input variations. The optimization objective incorporates a regularization term for the Jacobian norm: where $\|J_A(x)\|_F^2$ represents the squared Frobenius norm of the Jacobian matrix of the encoder function A with respect to the input x, and λ controls the strength of the regularization[4, 55].

- **Feed-Forward Autoencoders:** The structure of FFA includes an odd number of layers which ensures symmetry around the central layer. For instance, in a network handling inputs of dimension $n = 30$, the layers might be configured as follows: the first layer (n_1) has 30 neurons, the middle layer (n_2) reduces to 15 neurons, and the final layer (n_3) expands back to 30 neurons[55]. This illustration depicts the encoder-decoder structure inherent to FFAs.

    i. **Encoder:** The sequence of layers from the input to the bottleneck, progressively compressing the data into a lower-dimensional representation. Mathematically represented as $h_i = g(x_i)$, where g is the encoding function parameterized by the network's weights up to the bottleneck.

    ii. **Decoder:** The layers from the bottleneck back to the output, reconstructing the input data from the compressed form. The decoder is mathematically represented as $\hat{x}_i = f(h_i) = f(g(x_i))$, f being the decoding function.

Wasserstein Autoencoder (WAEs)

This is a variant of the autoencoder that aims to improve upon the Variational Autoencoder (VAE) by using the Wasserstein distance as a measure of similarity between the generated data distribution and the target data distribution[4]. Unlike VAEs, which minimize the Kullback-Leibler (KL) divergence to ensure the encoded latent space distribution approximates a prior distribution (often a Gaussian), WAEs focus on minimizing the Wasserstein distance, which can lead to more stable training and better quality of generated samples. Mathematically, WAE is represented as:

$$\text{minimize } \mathbb{E}_{P_X}[\text{loss}(x, x')] \tag{2.21}$$

where P_X is the distribution of the input data, x is an input sample, and x' is its reconstruction, $\text{loss}(x, x')$ is a reconstruction loss (e.g., MSE), ...

Deep Feature Consistent Variational Autoencoder (DFCVAE)

This introduces a novel loss function for optimizing autoencoders that focuses on the correlation between pixels rather than just the pixel-wise difference. Unlike traditional methods that measure direct pixel differences, this approach considers how pixels relate to each other. Variational Autoencoders (VAE) and Wasserstein Autoencoders (WAE) both aim to minimize reconstruction errors and a regularization term. However, VAE tries

to align each input's encoded distribution closely with a target distribution, leading to potential overlaps and reconstruction issues. WAE, on the other hand, matches the overall distribution of encoded inputs to the target, allowing for more distinct and spread-out latent representations, which enhances reconstruction quality[4]. Pretrained classification networks are often repurposed for tasks like transfer learning and style transfer, leveraging their learned features for new domains. Similarly, autoencoders can utilize these pretrained networks to define a more nuanced loss function. By comparing the original and reconstructed images' representations within these networks, a more sophisticated and meaningful measure of similarity is achieved, focusing on feature consistency rather than pixel-level accuracy[4].

Conditional Image Generation with Pixel-CNN Decoders (CIG-PCNND)

PixelCNN offers a unique approach to image generation by considering the spatial relationships between pixels. It generates images pixel by pixel, using a predefined order (e.g., top to bottom, left to right) and takes into account the local spatial context, which helps in producing more coherent and less blurred images. This method has evolved to incorporate Recurrent Neural Networks (RNNs) for capturing local statistics, maintaining the sequential pixel generation concept. Integrating PixelCNN as a decoder within autoencoders allows for generating images in a structured manner, where each pixel's generation considers both its predecessors and the encoded representation, leading to detailed and contextually accurate reconstructions[4].

Balancing Bias-Variance in Autoencoders

The design of an autoencoder must navigate the bias-variance trade-off: aiming for low reconstruction error (bias) while ensuring that the learned representations are generalizable and meaningful (variance). The choice of regularization technique plays a crucial role in achieving this balance, influencing both the quality of the data representation and the autoencoder's ability to reconstruct inputs accurately. Through these regularization strategies, autoencoders can be tailored to learn representations that capture the underlying structure of the data, facilitating a wide range of applications from compression to generative modeling[4].

2.7 Variational Autoencoder (VAE)

Variational Autoencoders, introduced by Kingma and Welling in 2019, extend the concept of traditional autoencoders by introducing a probabilistic twist. VAEs are designed to generate new data that is like the training data. Unlike standard autoencoders, which learn a deterministic function for encoding and decoding, VAEs model the encoding as a probability distribution over the latent space. This probabilistic approach allows VAEs to generate new data by sampling from the latent space distribution.

VAEs consist of two main components: the encoder, which maps the input data to a distribution over the latent space, and the decoder, which samples from this distribution to generate data that resembles the input data. The training process involves optimizing not only the reconstruction loss but also a regularization term that encourages the learned distribution to approximate a prior distribution, typically a Gaussian. This regularization term is crucial for ensuring that the latent space has good properties for data generation. [42, 71].

In variational autoencoders, VAEs, the core idea is to learn the parameters that best describe the data generation process by maximizing the marginal log-likelihood of the data. However, directly maximizing this likelihood is challenging, so VAEs focus on maximizing a lower bound on this likelihood, known as the variational lower bound. This bound is a function of the encoder and decoder parameters and is maximized to improve the approximation of the posterior distribution. The variational lower bound combines two terms: one penalizing the

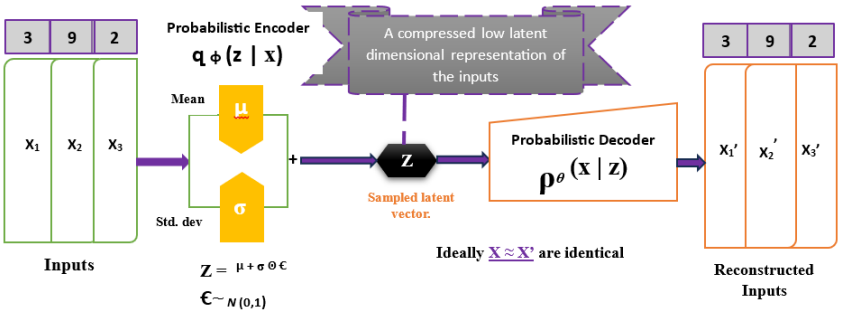


Figure 2.9: Simple Variational Autoencoder Architecture

difference between the approximate and true posterior distributions (using Kullback-Leibler divergence) and another based on the expectation of the log-likelihood of the data given the latent variables. Maximizing this lower bound effectively trains the VAE, making it a powerful tool for generative modeling. In practice, VAEs use stochastic gradient optimization and a technique known as the reparameterization trick to efficiently estimate gradients and update model parameters. This approach allows VAEs to learn complex data distributions and generate new data points that are like the observed data, making them highly useful for tasks like image generation and feature extraction.

Instead of encoding an input x to a fixed latent representation z, the encoder in a VAE maps x to a distribution over the latent space characterized by parameters (e.g., mean μ and variance σ^2) of a Gaussian distribution.
Encoder:

$$q_\phi(z|x) = \mathcal{N}(z; \mu(x), \sigma^2(x)) \tag{2.22}$$

Here, $q_\phi(z|x)$ represents the approximate posterior distribution of the latent variable z given an input x, parameterized by ϕ, with $\mu(x)$ and $\sigma^2(x)$ being the mean and variance of z.
Reparameterization Trick: To enable backpropagation, VAEs use the reparameterization trick for sampling z from the distribution:

$$z = \mu(x) + \sigma(x) \cdot \epsilon \tag{2.23}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is a noise term.
Decoder: The decoder part remains similar to AEs, aiming to reconstruct x from z:

$$\hat{x} = g_\theta(z) \tag{2.24}$$

The objective function of a VAE includes two terms: a reconstruction loss (similar to AEs) and a regularization term that encourages the learned distribution $q_\phi(z|x)$ to be close to the prior distribution $p(z)$, typically assumed to be a standard Gaussian $\mathcal{N}(0, I)$. This is often expressed using the Kullback-Leibler (KL) divergence:

$$L(x, \hat{x}) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL(q_\phi(z|x)||p(z)) \tag{2.25}$$

Key Mathematical Difference
The key mathematical difference between AE and VAE lies in the encoding process and the objective function. AEs learn a deterministic mapping to a fixed latent representation, focusing solely on minimizing the reconstruction error. In contrast, VAEs model the latent space probabilistically, optimizing not just for reconstruction but also for the distribution of the latent variables to approximate a prior distribution, incorporating the KL divergence into their loss function to achieve this balance.

Table 2.10 illustrates the fundamental differences between Autoencoders and Variational Autoencoders, focusing on their encoding processes, the nature of the latent space, reparameterization techniques, decoding processes, objective functions, and their key focuses.

Feature	Autoencoder (AE)	Variational Autoencoder (VAE)
Encoder	Maps input x to a latent representation z: $z=f(x)$	Maps input x to parameters of a probability distribution over z: (q_\phi (z)
Latent Space	Fixed latent representation z	Probabilistic latent space characterized by distribution parameters (μ, $\sigma 2$)
Reparameterization	Not applicable	Uses the reparameterization trick for sampling: $z=\mu(x)+\sigma(x)\cdot\epsilon$, $\epsilon\sim$N (0, I)
Decoder	Maps z back to reconstruction $x^$: $x^=g(z)$	Like AE, reconstructs x from z: $x^= g_\theta^{(z)}$
Objective Function	Minimize reconstruction error: $L(x, x^) = \|x-x^\|2$	Minimize combined loss: Reconstruction loss + KL divergence: (-\math bb {E} {q\phi (z)
Key Focus	Reconstruction accuracy	Reconstruction accuracy + Probabilistic modelling of latent space

Figure 2.10: Key Differences Between Autoencoder and Variational Autoencoder

2.7.1 Activation Function of the Output Layer

ReLU

- Formula: $\text{ReLU}(x) = \max(0, x)$
- Use Case: Suitable when inputs are positive. Not ideal for negative inputs as it outputs zero for them.

Sigmoid

- Formula: $\sigma(x) = \frac{1}{1+e^{-x}}$
- Use Case: Best for inputs normalized between 0 and 1. Commonly used for binary classification problems or when output needs to be in the $[0, 1]$ range.

2.7.2 Loss Function

Mean Square Error (MSE)

- Formula: $L_{\text{MSE}} = \frac{1}{M}\sum(x_i - \tilde{x}_i)^2$
- Application: Universal, works well regardless of the activation function or input normalization. Ideal for regression problems.

Binary Cross-Entropy (BCE)

- Formula: $L_{\text{CE}} = -\frac{1}{M}\sum(x_{j,i}\log(\tilde{x}_{j,i}) + (1 - x_{j,i})\log(1 - \tilde{x}_{j,i}))$
- Application: Suitable when the output layer uses a sigmoid function and inputs are normalized between 0 and 1. Effective for classification.

2.7.3 Reconstruction Error

- Common Metric: Mean Squared Error (MSE)
- Formula: $\text{RE} \approx \text{MSE} = \frac{1}{M}\sum(x_i - \tilde{x}_i)^2$
- Purpose: Measures the difference between original and reconstructed inputs. A lower RE indicates better reconstruction quality.

This summary encapsulates the essential aspects of activation functions, loss functions, and the concept of reconstruction error in the context of autoencoders, providing a clear understanding of when and why to use specific functions and how to evaluate autoencoder performance.

Type of Autoencoder	Key Characteristics	Objective/Use
Standard Autoencoder	Simple encoder-decoder architecture. Learns to compress and reconstruct input data.	Basic feature learning and compression.
Variational Autoencoder (VAE)	Introduces a probabilistic graphical model with the encoder outputting parameters of a distribution.	Generative modelling, capable of generating new data points.
Sparse Autoencoder	Adds sparsity constraint on the hidden layers to induce a small number of active neurons.	Feature selection, robustness to noise.
Denoising Autoencoder	Adds noise to input data and learns to recover original data.	Robust feature learning and error correction.
Contractive Autoencoder	Adds penalty on the sensitivity of the hidden representation to the input. Focuses on stable feature extraction.	Robust feature extraction against small input variations.
Feed Forward Autoencoder (FFA)	Utilizes dense layers, typically with a symmetrical architecture around a bottleneck layer.	Efficient data compression and straightforward feature learning.
Wasserstein Autoencoder	Optimizes a different cost function to improve upon VAEs. Aims for a more stable training process and better sample quality.	Generating high-quality samples and improving stability in training generative models.
Deep Feature Consistent Variational Autoencoder	Uses a different loss function that considers the correlation between pixels, often leveraging pretrained networks for loss computation.	Improved visual fidelity in reconstruction, maintaining deep feature consistency.
Conditional Image Generation with Pixel CNN Decoders	Combines autoencoders with Pixel CNN decoders to generate images sequentially, considering local spatial statistics.	High-quality image generation with detailed texture and coherence.

Figure 2.11: Key Differences Between Autoencoder and Variational Autoencoder

Table 2.11 summarizes the distinct features and primary applications of each type of autoencoder, highlighting their versatility in various tasks such as dimensionality reduction, feature learning, data denoising, and generative modeling.

2.8 Autoencoders and Generative Adversarial Networks

Variational Autoencoders (VAEs) typically use Mean Squared Error (MSE) for training, which can lead to slightly blurred images. However, they allow for inference over latent variables, giving some control over the generated output[4]. On the other hand, Generative Adversarial Networks (GANs) consist of a generator that creates new samples and a discriminator that differentiates between real and generated samples. The training process, which involves a competitive loss function, enhances the quality of generated data but at the cost of control over the output. Several approaches have been explored to combine the benefits of VAEs and GANs. For instance, Adversarial Autoencoders replace the KL-divergence in VAEs with a discriminator that differentiates between the prior and posterior distributions. Other variations modify the reconstruction loss in VAEs with a discriminator, blending the decoder with the generator, or integrate the GAN's discriminator with an encoder, facilitating inference in the latent space. These hybrid models aim to leverage the strengths of both architectures for improved data generation and inference capabilities[4].

2.9 Application Domains of Autoencoders

Autoencoders, particularly Variational Autoencoders (VAEs), serve as powerful generative models, enabling the generation of new data samples by learning a probabilistic distribution of the data. They find applications across various domains, from enhancing classification results to clustering, anomaly detection, recommendation systems, and dimensionality reduction. Here's a simplified overview of their applications:

Generative Models

Variational Autoencoders (VAEs) can generate new data samples by sampling from a learned probabilistic distribution. This capability allows for the creation of new, meaningful samples once the model is trained, as demonstrated with datasets like Tabular data, images or MNIST[4].
It stands out in the generative models domain for their ability to learn complex data distributions and generate new, unseen data points that mimic the original dataset. This is particularly useful in fields like drug discovery, where VAEs can generate novel molecular structures for potential pharmaceuticals, or in content creation, where they can produce diverse and realistic images, music, or text. For example, VAEs have been used to generate realistic human faces, artwork, and even to interpolate between different musical genres, demonstrating their versatility and power in creating high-quality, diverse outputs from learned data distributions.

Classification Enhancement

Autoencoders, by learning efficient data representations, can significantly enhance classification tasks, especially in semi-supervised learning scenarios. This is evident in image recognition, where autoencoders preprocess images to extract salient features, which are then used by classification models to improve accuracy[4]. For instance, in facial recognition systems, autoencoders can help isolate features such as edges, shapes, and textures that are crucial for identifying individuals, thereby improving the performance of the classifier even with limited labeled data.

Clustering

By compressing data into a lower-dimensional latent space, autoencoders simplify the clustering of complex datasets by highlighting inherent groupings within the data[4]. This approach has been beneficial in customer segmentation, where businesses can cluster customers based on purchasing behavior, preferences, and other characteristics to tailor marketing strategies effectively. In genomics, autoencoders have facilitated the clustering of genetic data, helping researchers identify patterns and similarities in genetic expressions that might not be apparent in the high-dimensional original data.

Anomaly Detection

Autoencoders excel in anomaly detection by learning to reconstruct normal data while failing to do so accurately for anomalies, resulting in high reconstruction errors for the latter[4]. This property is invaluable in cybersecurity, where autoencoders can detect unusual patterns in network traffic that may indicate a security breach as fraud detection. Similarly, in manufacturing, autoencoders monitor equipment data to identify signs of future failures, allowing for preventive maintenance and reducing downtime.

Recommendation Systems

In recommendation systems, autoencoders, such as AutoRec, have shown promise by compressing user-item interaction matrices into dense representations that capture user preferences and item characteristics[4]. This approach has been applied in online retail to suggest products to users based on their browsing and purchase history, significantly improving user engagement and sales. Similarly, in content streaming services, autoencoders help recommend movies, TV shows, and music tailored to individual tastes, enhancing user experience by providing personalized content recommendations.

Dimensionality Reduction

Autoencoders are a nonlinear alternative to PCA for dimensionality reduction, capable of capturing complex data structures in a lower-dimensional space. This application is crucial for dealing with high-dimensional data, such as images or text, and mitigating the curse of dimensionality[4].
Each application leverages the unique ability of autoencoders to learn efficient representations of data, whether for generating new samples, enhancing classification and clustering, detecting anomalies, making recommendations, or reducing dimensionality. These applications underscore the versatility and utility of autoencoders in various machine learning and data processing tasks.

2.10 Privacy

Privacy encompasses the protection of personal data from unauthorized access, disclosure, or misuse, ensuring that individuals' confidentiality and autonomy are respected. Data privacy is a critical aspect of information technology and digital services, governed by laws and regulations like the General Data Protection Regulation (GDPR) in the European Union, which sets strict guidelines for data handling and grants individuals specific rights over their data[32].

Data utility, on the other hand, refers to the usefulness or value that data provides, especially when used for decision-making, analysis, and other processes. In the context of privacy, there is often a balance or trade-off to be managed between protecting individuals' privacy and maximizing the utility of data. For example, anonymizing data can protect privacy but may reduce its utility for certain types of analysis. The challenge lies in implementing privacy-preserving techniques that minimize the impact on data utility, such as differential privacy, which adds noise to data in a way that protects individual privacy while still allowing for meaningful analysis[32].

Regulatory Landscapes: GDPR and HIPAA

In the era of digital transformation, the protection of personal data has risen to the forefront of global priorities. Two critical regulatory frameworks that guide data privacy and security practices today are the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States[1, 10, 32].

The General Data Protection Regulation (GDPR)

Adopted on April 14, 2016, and effective from May 25, 2018, the GDPR represents a comprehensive data protection law in the EU. It aims to give individuals control over their personal data while simplifying the regulatory environment for international business by unifying the regulation within the EU. The GDPR impacts any organization, regardless of

location, that processes personal data of EU residents. It emphasizes principles such as consent of the data subject, data minimization, and the right to erasure, also known as the right to be forgotten. Violations of GDPR can lead to significant fines, up to €20 million or 4% of the annual worldwide turnover of the preceding financial year, whichever is higher[10]. For further information on GDPR, please refer to the official website of the European Commission: GDPR | European Commission.

The Health Insurance Portability and Accountability Act (HIPAA)

Enacted on August 21, 1996, HIPAA is a United States legislation that provides data privacy and security provisions for safeguarding medical information. The primary goal of HIPAA is to make it easier for people to keep health insurance, protect the confidentiality and security of healthcare information, and help the healthcare industry control administrative costs. HIPAA applies to covered entities and their business associates that handle protected health information (PHI). It establishes national standards for the protection of health information, as well as civil and criminal penalties for violations[1]. For more insights into HIPAA, visit the official U.S. Department of Health & Human Services website: HIPAA for Professionals | HHS.gov.

Implications for Data Privacy and Synthetic Data Generation

The advent of synthetic data generation offers promising avenues for leveraging extensive datasets while adhering to GDPR and HIPAA requirements. However, ensuring that synthetic data generation processes and the resultant datasets comply with these regulations is paramount. The use of tools like Anonymeter to assess privacy risks in synthetic datasets emerges as a crucial step in aligning synthetic data initiatives with GDPR and HIPAA, ensuring that the benefits of synthetic data are harnessed responsibly and ethically[1].

2.10.1 Understanding Anonymeter: A Comprehensive Tool for Privacy Risk Assessment in Synthetic Datasets

In the quest for balancing data utility and privacy, synthetic datasets have emerged as a promising solution, enabling the exploration and analysis of data while preserving individual privacy. However, the generation of synthetic data introduces its own set of challenges, particularly concerning the assessment and mitigation of privacy risks. This is where Anonymeter, developed by Anonos, becomes indispensable[22].

The Essence of Anonymeter

Anonymeter serves as an advanced tool designed to scrutinize privacy risks in synthetic datasets meticulously. Recognized by CNIL, France's data protection authority, for its capabilities, Anonymeter plays a crucial role in ensuring synthetic datasets are effectively anonymized, addressing both anonymization and pseudonymization concerns[22].

Operational Mechanics

Anonymeter stands out by evaluating synthetic datasets through the simulation of various privacy attacks, focusing on the potential for re-identification. It operationalizes this by[22]:

- **The Attack Phase:** Making educated guesses about individual records within a synthetic dataset.
- **The Evaluation Phase:** Comparing these educated guesses against the original dataset to ascertain the accuracy of the privacy attacks.

- **The Risk Estimation Phase:** Determining the actual privacy risks by conducting the attack on a control dataset not involved in the synthetic data's generation.

Key Attributes of Anonymeter

- **Comprehensive Risk Evaluation:** Anonymeter assesses risks pertaining to singling out, linkability, and inference, offering a detailed perspective on privacy concerns.
- **Balancing Utility and Privacy:** It provides insights into the trade-offs between the informational value of synthetic data and its privacy implications.
- **Ease of Use:** Anonymeter is designed with user accessibility in mind, requiring minimal technical expertise to implement.

Anonymeter's methodology encompasses several key components, each targeting a specific aspect of privacy risk[22]:

1. **Singling-Out Risk Assessment:** This component evaluates the risk associated with the possibility of identifying an individual from a dataset. It includes:
    - *Univariate Analysis:* Assesses the risk of identification from a single attribute.
    - *Multivariate Analysis:* Evaluates the risk when multiple attributes are combined, providing a more nuanced analysis of privacy risk.
2. **Linkability Risk Assessment:** Measures the risk that two or more records from different datasets can be linked to the same individual, potentially revealing sensitive information.
3. **Inference Risk Assessment:** Assesses the likelihood that sensitive information can be inferred from an anonymized or synthetic dataset, emphasizing the need for robust anonymization techniques.

Practical Application

For organizations aiming to navigate the complexities of data privacy while leveraging the benefits of synthetic datasets, Anonymeter offers a direct and efficient solution. This tool simplifies the privacy risk assessment process, making it accessible for a broad range of users, from data scientists to privacy professionals.

1. **Installation and Setup:** The first step in utilizing Anonymeter involves accessing its source code on GitHub. Here, users can find comprehensive instructions for cloning the repository and setting up the environment required to run Anonymeter. This process is designed to be straightforward, ensuring that even those with limited technical expertise can successfully install and begin using the tool.
2. **Configuration:** After installation, the next critical step is the configuration of Anonymeter to suit specific evaluation needs. This phase involves defining the auxiliary information that is known outside of the dataset and specifying the target attributes for privacy risk evaluation. Such configuration allows Anonymeter to simulate realistic attack scenarios, accurately measuring the potential risks of re-identification or information inference. Through this targeted approach, Anonymeter can provide insights into the specific areas of vulnerability within synthetic datasets, guiding users in implementing effective mitigation strategies.

For those interested in integrating Anonymeter into their synthetic data pipeline, detailed documentation and the necessary resources can be found at the Anonymeter GitHub repository: https://github.com/statice/anonymeter. This GitHub page serves as a hub for all information related to Anonymeter, including updates, user guides, and community

support. By offering open access to the tool, Anonos encourages ongoing collaboration and development within the privacy technology community, ensuring that Anonymeter remains a cutting-edge solution for privacy risk assessment in the evolving landscape of synthetic data.

Why Choose Anonymeter?

Anonymeter's development as an open-source tool underlines Anonos's commitment to enhancing privacy technologies' accessibility. It is crafted to be adaptable, ensuring it remains relevant amid evolving privacy regulations and research advancements. The Anonymeter framework plays a pivotal role in the field of synthetic data by providing a structured approach to evaluate and mitigate privacy risks. By assessing singling-out, linkability, and inference risks, Anonymeter helps researchers and practitioners balance the trade-offs between maintaining data utility and ensuring privacy. This is particularly relevant in the era of big data and machine learning, where the use of synthetic data is becoming increasingly prevalent.

Looking Ahead

Anonos envisions expanding Anonymeter's functionalities to include new privacy metrics and support for various data types and de-identification methods. This forward-looking approach ensures Anonymeter remains at the forefront of privacy protection technology, aiding organizations in navigating the complex landscape of synthetic data privacy with confidence.

Reference to Anonymeter

For further details on Anonymeter and its methodology, readers are referred to the official documentation and publications by the creators of Anonymeter. This source provides comprehensive insights into the framework's development, application, and impact on privacy-preserving data analysis.

Note: For specific details and technical descriptions of the Anonymeter framework, please refer to the work of Elise Devaux [22].

2.10.2 Autoencoders (AE) and Data Privacy Preservation

Autoencoders operate by compressing input data into a lower-dimensional representation (latent space) and subsequently reconstructing the original data from this compressed form. This process is pivotal for privacy preservation for several reasons. Firstly, the encoding phase abstracts away from the specifics of individual data points, focusing instead on capturing the underlying distribution and patterns within the dataset. By learning a generalized representation, AEs ensure that the synthetic data generated does not replicate any specific individual's data. Instead, the reconstructed data reflects the aggregate characteristics of the input data, thereby safeguarding individual privacy.

2.10.3 Variational Autoencoders (VAE) and Enhanced Data Privacy

Variational Autoencoders take the privacy-preserving capabilities of AEs further by introducing a probabilistic twist to the encoding and decoding process. VAEs generate a latent space that is not just a compressed representation but a probabilistic distribution of possible representations. The key component here is the sampling layer, which introduces randomness into the generation of the latent variables. This randomness means that any synthetic data point generated by a VAE is essentially a sample from the learned distribution, further distancing the synthetic instance from any real individual's data. The probabilistic nature of VAEs inherently embeds a layer of uncertainty, making it significantly more challenging to trace back to any specific real-world data point, thus enhancing privacy.

2.10.4 The Privacy-Preserving Mechanism in AE and VAE

The privacy-preserving aspect of AE and VAE models is not an add-on but a fundamental characteristic of how these models operate. By learning to encode the data into a new space (latent space for AE and probabilistic latent space for VAE) and then decoding it to generate new data points, these models ensure that the output reflects the collective features of the dataset rather than any individual's data. This methodology inherently protects patient data privacy and individual data confidentiality by design. The generated synthetic data maintains the utility for research and development purposes, such as training machine learning models or conducting statistical analyses, without compromising the privacy of the individuals whose data were used to train the generative models.

The application of AE and VAE for synthetic data generation inherently employs data privacy techniques through their operational mechanisms. The process of encoding, transforming, and decoding the data in AEs, coupled with the probabilistic generation in VAEs, ensures that synthetic data generation is naturally aligned with privacy preservation goals. This inherent privacy-preserving feature makes AE and VAE powerful tools for generating synthetic datasets that are both useful and respectful of individual privacy, offering a robust approach to handling sensitive data in healthcare and beyond. The approaches of leveraging AE and VAE are generally referred to as differential privacy when specifically designed to minimize the risk of re-identification, or more broadly, it can be considered as part of synthetic data generation techniques for privacy preservation.

2.11 Related Works

In the context of synthetic data generation, [51] study highlights the application of GANs for creating synthetic census microdata, revealing the complexities and risks of using GANs, especially for mixed-type data like healthcare records. This work underscores the need for integrating Autoencoders (AEs) and Variational Autoencoders (VAEs), subsets of GANs to tackle these challenges effectively.

Building on this, [62] research, "Exploring the Value of GANs for Synthetic Tabular Data Generation in Healthcare with a Focus on Data Quality, Augmentation, and Privacy," delves into the potential of GANs to mimic the complexity of real healthcare datasets while safeguarding privacy. Pedersen's findings, particularly with CTGAN and CopulaGAN models, resonate with the thesis's aim to enhance synthetic data's security and utility.

In the paper "Generation and Evaluation of Tabular Data in Different Domains Using GANs" [54], the authors explore the capabilities of Generative Adversarial Networks (GANs) in creating and assessing structured, high-dimensional tabular data across a variety of fields. This research highlights the flexibility of GANs in generating synthetic datasets that closely mimic real data, suggesting the integration of Autoencoders (AEs) and Variational Autoencoders (VAEs) to enhance data security and privacy.

The study presented in [47] introduces an innovative method that incorporates expert knowledge into the synthetic data generation process. By addressing common challenges such as bias and overfitting in GAN models, this approach demonstrates its alignment with the goals of this thesis, particularly in the context of generating more accurate and secure synthetic data for healthcare applications.

In the notable study, by Stadler, Oprisanu, and Troncoso (2022) critically evaluate the effectiveness of synthetic data as a means to preserve privacy in the context of data publishing. Their work, presented at the 31st USENIX Security Symposium, addresses a fundamental question in the field of data privacy: Can synthetic data, generated from state-of-the-art generative models, adequately protect against inference attacks while maintaining data utility?

[74]

Their research, titled "Synthetic Data – Anonymisation Groundhog Day," challenges the prevailing notion that synthetic data offers a foolproof solution to privacy concerns associated with traditional anonymisation techniques. The authors provide a quantitative assessment that demonstrates synthetic data often fails to offer a better balance between privacy protection and data utility compared to traditional methods. This finding is significant as it empirically contests the idea that synthetic data can serve as a silver bullet for privacy-preserving data publishing [74].

Stadler, Oprisanu, and Troncoso's study reveals that the privacy-utility tradeoff in synthetic data publishing is unpredictably variable. They argue that unlike traditional anonymisation, where the effects and limitations are somewhat predictable, synthetic data does not consistently guarantee that all sensitive information is masked or that the statistical properties of the original data are preserved. This unpredictability can lead to significant variations in privacy gains and unforeseen losses in utility [74].

The implications of their findings are critical for researchers and practitioners in data privacy, as it underscores the necessity of a cautious approach to using synthetic data. Their work suggests that relying solely on synthetic data for privacy preservation could be misguided without thorough validation and understanding of the generative models' capabilities and limitations[74].

Utilization of Autoencoders and Variational Autoencoders in Healthcare and Other Domains' Tabular Data

In the rapidly evolving field of healthcare, the advent of machine learning technologies has opened new avenues for enhancing patient care, diagnosis, and treatment processes. Among these technologies, autoencoders have emerged as a powerful tool for dealing with complex healthcare data. Autoencoders, through their unique ability to learn efficient representations and features from vast amounts of unlabelled data, have found applications ranging from anomaly detection in patient records to the generation of synthetic data for research purposes. This subsection of chapter 2 demonstrates the diverse applications of autoencoders in healthcare, highlighting their potential to transform medical data analysis, enhance privacy and security, and contribute to the advancement of personalized medicine. By exploring various studies and implementations, we aim to provide a comprehensive overview of how autoencoders are being utilized to address some of the most pressing challenges in healthcare today.

Synthetic Electronic Health Records Generated with Variational Graph Autoencoders

Nikolentzos et al.'s (2023), "Synthetic electronic health records generated with variational graph autoencoders," introduces an innovative method for creating synthetic patient trajectories from electronic health records (EHRs), aiming to overcome privacy concerns in healthcare data usage. Leveraging Variational Graph Autoencoders (VGAEs), the research showcases the ability to generate synthetic data that is both clinically realistic and privacy-compliant, effectively capturing the intricate time-dependencies and correlations found in patient data[57]. This approach not only ensures the generation of large, complex graphs that mirror the statistical characteristics of actual health records without risking patient privacy but also marks a significant step forward in enabling the secure sharing and utilization of healthcare data. By providing a novel solution to the data accessibility challenge in healthcare, Nikolentzos et al.'s work holds the potential to significantly advance medical research and the implementation of AI technologies in healthcare settings, all while adhering to strict privacy standards.

Synthesising Multi-Modal Minority Samples for Tabular Data

The paper "Synthesising Multi-Modal Minority Samples for Tabular Data" by Sajad Darabi and Yotam Elor introduces a novel framework, Tabular AutoEncoder Interpolator (TAEI), aimed at addressing the challenge of imbalanced binary classification in machine learning, particularly for tabular datasets that include both continuous and categorical (discrete) features. Recognizing the limitations of traditional oversampling methods like SMOTE in handling multi-modal data, the authors propose leveraging autoencoders to map samples into a dense continuous latent space, where interpolation can effectively generate high-quality synthetic minority samples. This process not only enhances the representation of minority classes but also significantly improves the prediction accuracy in downstream binary classification tasks[20]. Through extensive experimentation across 27 real-world datasets, the framework demonstrated superior performance in generating synthetic data, outperforming existing methods including GAN-based approaches like CTGAN and TGAN. The study introduces new metrics for directly assessing the quality of generated data, underscoring the framework's ability to produce realistic synthetic samples and its potential for broad application in machine learning models dealing with imbalanced data.

EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders

In the paper "Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders" by Siddharth Biswal et al. introduces EVA, a deep generative model designed to synthesize realistic sequences of Electronic Health Records (EHR) encounters, addressing the critical need for large, realistic EHR datasets for healthcare research while ensuring patient privacy. Leveraging a combination of stochastic gradient Markov Chain Monte Carlo and amortized variational inference, EVA efficiently generates EHR sequences that reflect individual patient differences and can be conditioned on specific disease conditions, enabling targeted disease-specific studies. Evaluated on extensive real-world EHR data from over 250,000 patients, the model demonstrates its ability to produce sequences that clinicians find realistic, with predictive models trained on synthetic data performing comparably to those trained on actual EHRs[6]. Furthermore, augmenting real data with synthetic EHRs generated by EVA improves predictive performance, showcasing EVA's potential to significantly advance healthcare research while safeguarding patient privacy.

Transfer Learning for Tabular Data

In the paper "Transfer Learning for Tabular Data" by Leonid Joffe, a novel deep learning architecture is introduced, aiming to overcome the limitations of traditional models that are confined to specific table formats in tabular data [39]. This architecture, inspired by the universal applicability of computer vision models, is designed to capture useful patterns across arbitrary tables by training on randomly sampled subsets of features processed by a convolutional network. This approach enables the model to learn feature interactions within the table, producing embeddings that are transferable and enhance the performance of classifiers across various machine learning benchmark datasets. The paper demonstrates that these embeddings, when used as additional features, significantly improve classification accuracy, suggesting the potential of transfer learning in the realm of tabular data. This research opens new avenues for applying deep learning to tabular datasets, where the model's ability to abstract and generalize from feature interactions can lead to more versatile and effective machine learning applications.

Advancements in Biomedical Machine Learning

Recent advancements in machine learning, particularly through the application of autoencoders (AEs) and variational autoencoders (VAEs), have shown promising results in the biomedical field, offering novel approaches to the diagnosis of rare diseases (RDs) and the handling of complex omics data. Pratella et al. provide a comprehensive overview of how these unsupervised learning models are being utilized to enhance the understanding and diagnostic processes of RDs, leveraging the vast and intricate datasets generated by high-throughput sequencing technologies. These technologies, including AEs and VAEs, have emerged as powerful tools for deciphering the complex, high-dimensional data characteristic of modern biomedical research. By compressing data into a more manageable latent space, AEs and VAEs facilitate the extraction of meaningful patterns and features that are often obscured in raw datasets, demonstrating their potential to provide insights into cancer, bacterial infections, and the physiological states of healthy tissues[65]. The principles underlying the application of AEs and VAEs in RD diagnosis and omics data analysis bear significant relevance to the challenges of securing synthetic healthcare data. In generating synthetic datasets that maintain the statistical properties of original data while ensuring privacy, the ability of AEs and VAEs to capture and encode essential data characteristics becomes invaluable. This process mirrors the compression and reconstruction mechanism of AEs, where sensitive data elements are abstracted into a latent representation, effectively obfuscating individual data points to protect patient privacy. Moreover, the adaptability of AEs and VAEs to handle various data complexities and their application in denoising and data integration offer important lessons for enhancing data security. By applying these models to generate synthetic healthcare data, researchers can ensure that the resulting datasets are not only diverse and representative but also devoid of direct identifiers, thereby mitigating the risk of re-identification.

Advancements in Personalized Cancer Treatment through Machine Learning

According to Hongyuan Dong et al. (2021), the development of personalized cancer treatment strategies and the discovery of new anti-cancer drugs are significantly enhanced by leveraging Variational Autoencoders (VAEs) and Multi-Layer Perceptrons (MLPs). Their innovative approach, which combines gene expression data from cancer cell lines with molecular data of anti-cancer drugs, utilizes a novel GENEVAE model for gene expression data and a rectified Junction Tree Variational Autoencoder (JTVAE) for drug molecular data. This methodology not only facilitates the prediction of drug responses with remarkable accuracy, as evidenced by high R^2 scores, but also demonstrates the potential for generating novel drug compounds effective against specific cancer cell lines. Their work marks a significant advancement in the application of machine learning techniques in oncology, offering promising avenues for accelerating the discovery of effective cancer treatments and advancing personalized medicine[24].

Comparative Analysis of Computational Models in Oncology

In exploring the efficacy of various computational models in predicting anti-cancer drug responses, Dong et al. (2021) presents a comprehensive comparison of six models combining gene expression data and drug molecular structures with different analytical approaches [24]. The performance of these models on both breast cancer and pan-cancer datasets is summarized, showcasing the predictive accuracy through R^2 scores and RMSE metrics. The results underscore the superior performance of models that integrate Variational Autoencoders (VAEs) with Multi-Layer Perceptrons (MLPs), particularly when applied to a broad spectrum of cancer types.

Comparative Performance of Predictive Models in Oncology

Table 1 below illustrates the significant impact of incorporating Variational Autoencoders (VAEs) into predictive models. Notably, the CGC + VAE + MLP model achieves the highest R^2 scores, indicating strong predictive accuracy across both breast and pan-cancer datasets. This evidence supports the potential of VAE-enhanced models in improving the precision of drug response predictions, thereby contributing to the development of more effective, personalized cancer treatments.

Table 2.2: Performance of the 6 proposed models on breast and pan cancer datasets

Models	Cancer Type	R^2 Score	RMSE
CGC + SVR	Breast	0.658	1.582
CGC + VAE + SVR	Breast	0.692	1.491
CGC + MLP	Breast	0.822	1.133
RAW + VAE + MLP	Breast	0.805	1.163
CGC + VAE + MLP	Breast	0.830	1.130
CGC + VAE + MLP	Pan cancer	0.845	1.080

Source: Adapted from Dong et al. (2021).

Causal Recurrent Variational Autoencoder for Medical Time Series Generation

In the paper *"Causal Recurrent Variational Autoencoder for Medical Time Series Generation"* by Hongming Li, Shujian Yu, and Jose Principe, the authors introduce the Causal Recurrent Variational Autoencoder (CR-VAE), a novel generative model designed to learn and incorporate Granger causality into the generation of multivariate time series data. Unlike traditional models, CR-VAE features a multi-head decoder to handle the generation of each time series dimension while learning a sparse adjacency matrix that encodes causal relationships, ensuring the data generation process adheres to the principles of Granger causality[49]. The model's effectiveness is demonstrated through experiments on synthetic data and real-world medical datasets, including EEG and fMRI signals, where it outperforms state-of-the-art time series generative models in both the quality of synthetic data generation and the accuracy of causal discovery. This advancement highlights CR-VAE's potential to enhance transparency in the generative process and its applicability in medical and healthcare domains where understanding causal relationships is crucial.

Innovative Drug Discovery Against COVID-19

In the face of the global COVID-19 pandemic, the innovative work by Cheng et al. (2021) introduces the Genetic Constrained Graph Variational Autoencoder (GCGVAE), a groundbreaking model designed for the rapid discovery of therapeutic drugs against SARS-CoV-2. Trained on protein structure data from a variety of viruses, including SARS, HIV, Hep3, and MERS, the GCGVAE model employs advanced optimization algorithms such as valency masking and genetic algorithms to fine-tune the generation of potential drug moleculescheng2021genetic[13]. This approach not only accelerates the identification of viable drug candidates but also ensures the structural feasibility of these molecules, showcasing the model's ability to produce compounds with high binding affinity to the viral protease of SARS-CoV-2. The simulation results presented by Cheng et al. demonstrate that the molecules generated by GCGVAE significantly outperform existing drugs in terms of effectiveness, underscoring the model's potential to contribute meaningfully to the fight against COVID-19. Moreover, the GCGVAE model's versatility extends beyond COVID-19, offering a promising framework for drug discovery against other viral pathogens. By automating the selection of active molecules from extensive databases and generating structurally viable drug candidates,

the model addresses critical challenges in the drug development process, reducing both time and cost. Cheng et al.'s work exemplifies the power of computational models in pandemic response efforts, providing a scalable and efficient solution for the rapid development of therapeutic agents. The success of the GCGVAE model in generating superior drug candidates for COVID-19 treatment highlights its potential as a transformative tool in the field of drug discovery, paving the way for its application in addressing future global health crises[13].

Exploring Factor Structures with VAEs in Personality Research

Huang and Zhang (2022), in their pioneering study, leverage the Variational Autoencoder (VAE) to explore factor structures within personality research, challenging the traditional Linear Factor Analysis (LFA) approach. Through meticulous analysis of the International Personality Item Pool (IPIP) Big 5 and HEXACO datasets, the authors demonstrate VAE's superior capability in identifying stable factor structures, even as the number of assumed latent factors increases. Unlike LFA, which tends to fractionate factors into smaller, unstable components, VAE consistently identifies a more nuanced factor structure, suggesting its potential to exhaustively explore factor structures in a single process. This study not only highlights VAE's effectiveness in handling complex, non-linear associations between latent factors and personality variables but also underscores its limitations, such as the need for large sample sizes to ensure model convergence. Huang and Zhang's work opens new avenues for personality model construction, emphasizing the importance of incorporating non-linear analytical tools like VAE in psychological research[37].

PepVAE: A VAE Framework for Antimicrobial Peptide Generation

In the innovative study "PepVAE: Variational Autoencoder Framework for Antimicrobial Peptide Generation and Activity Prediction" by Dean et al. (2021), the authors introduce a groundbreaking approach to antimicrobial peptide (AMP) discovery using a variational autoencoder (VAE). This framework, PepVAE, is adept at generating novel AMP sequences and predicting their antimicrobial activity solely based on sequence and experimental data. By encoding AMP sequences into a latent space, PepVAE facilitates the generation of new AMPs with desired properties through controlled sampling from specific latent space regions. The study's validation process, involving experimental minimum inhibitory concentration (MIC) assays against pathogens like E. coli, S. aureus, and P. aeruginosa, confirms the efficacy of the generated AMPs. This method represents a significant leap forward in the quest for new antimicrobials, offering a rapid, efficient, and low-cost tool for developing peptides with potent bactericidal activities, thereby addressing the urgent need for novel antimicrobials in the post-antibiotic era[21].

ECAAE: Accelerating Drug Discovery with Generative Architecture

Polykovskiy et al. (2021) introduced an innovative generative architecture, the Entangled Conditional Adversarial Autoencoder (ECAAE), aimed at accelerating the drug discovery process through computational means. This model can generate novel molecular structures with specified properties, such as activity against proteins, solubility, or synthetic feasibility. The authors applied ECAAE to generate a novel inhibitor targeting Janus kinase 3, a protein implicated in various autoimmune diseases, demonstrating the molecule's efficacy through in vitro testing[64]. This work underscores the potential of generative models to significantly reduce the time and cost associated with traditional drug discovery methods, offering a promising avenue for rapid development of therapeutic agents. The ECAAE model addresses disentanglement issues present in previous models by incorporating predictive and joint approaches, ensuring the conditional generation of complex molecular structures. Furthermore, the model's semi-supervised extension allows for the utilization of partially labeled datasets, enhancing its applicability in real-world scenarios where complete property

datasets are rare[64]. This advancement in generative modeling represents a significant step forward in the use of machine learning techniques for drug discovery, providing a versatile tool for generating compounds with desired properties and potentially transforming the pharmaceutical development landscape.

Demystifying VAEs in Synthetic Financial Data Generation

Wu et al. in their 2023 study, introduce a sensitivity-based approach to demystify the process by which Variational Autoencoders (VAEs) generate synthetic financial data. This method illuminates the "black box" nature of VAEs, offering insights into how specific input features affect the model's latent space and, consequently, the synthetic data produced. Tested on both simulated and real banking datasets from Kaggle, the research showcases the technique's ability to clarify and quantify the role of input features in synthetic data creation, enhancing transparency and efficiency in financial applications where data privacy is critical[81]. The study's findings are not limited to finance; they hold promise for sectors like healthcare and education, where generating privacy-compliant synthetic data is essential[81]. By identifying key features that influence data generation, the sensitivity-based method improves model efficiency and fosters a deeper understanding of deep learning models' inner workings. Wu et al.'s work contributes significantly to the field of explainable AI, advocating for greater transparency and trust in machine learning and highlighting the potential for broader application of such interpretative techniques in sensitive and regulated environments.

Medical Image Compression Based on Variational Autoencoder

In the study "Medical Image Compression Based on Variational Autoencoder" by Liu et al., a novel approach to medical image compression is introduced, utilizing variational autoencoders (VAEs) combined with residual network modules. This method addresses the challenge of efficiently compressing medical images amidst the "explosive" growth of medical data, constrained by limited network bandwidth and storage capacity. The algorithm optimizes both the compression rate and the distortion of reconstruction simultaneously, surpassing traditional compression techniques that struggle with dual optimization. By incorporating residual networks, the algorithm effectively minimizes information loss during compression, ensuring the preservation of critical medical details. Experimental results demonstrate superior performance in terms of lower distortion and better reconstruction effects compared to existing medical image compression algorithms, maintaining high-quality image reconstruction across various compression rates[53]. This advancement is significant for enhancing the storage and transmission efficiency of medical images, supporting the increasing demands of medical diagnostics and research.

Chapter 3

Methodology

This Chapter stands as the heart of this exploration, meticulously detailing the methodology that underpins our journey through the realms of synthetic data generation and evaluation. Within these pages, we embark on a deep dive into the innovative generative methodologies(autoencoder and variational autoencoder) that have been harnessed to create synthetic datasets that mirror the complexity and nuance of real-world data. Our narrative does not stop at creation; it extends into the rigorous assessment of the synthetic data's fidelity to its origins, employing a blend of established and novel evaluation metrics and privacy risk assessment leveraging Anonymeter.

The choice of Python version 3.8 as the scaffolding for this thesis is both a nod to practicality and a testament to the language's robust ecosystem. Python's widespread adoption and its rich repository of libraries have made it an indispensable tool in the data scientist's arsenal. The implementation of our generative methodologies, the orchestration of evaluation metrics, and the utilization of Anonymeter have all been meticulously documented and executed within this versatile programming environment.

As we navigate through the chapter, readers will gain insights into the meticulous processes involved in generating synthetic datasets, from conceptualization to realization. The methodologies, AE, and VAE detailed here are not merely academic exercises; they are practical tools refined and validated through empirical research. The datasets that emerge from these processes are then subjected to a rigorous evaluation, with the findings meticulously documented and made accessible for further inquiry. For a detailed methodologies, AE, and VAE look at the code and implementations used throughout these methodologies, visit our: GitHub Repository.

This chapter is not just a narrative of methods, metrics, and experimental designs; it is an invitation to explore the frontier of synthetic data generation and data privacy risk assessment techniques. It lays down the groundwork for a discourse that challenges conventional notions of data privacy and utility, inviting readers, researchers, and practitioners alike to engage with the material in a manner that is both critical and curious.

As you step into this chapter, be prepared to journey through the intricacies of synthetic data generation and evaluation, guided by the principles of transparency, rigor, and innovation. Welcome to the methodology of our thesis, where the foundation of our research is laid bare, ready to be built upon by inquisitive minds eager to push the boundaries of what is possible in the realm of synthetic data.

3.1 Dataset Description

In this study, three healthcare datasets were analyzed, each addressing a distinct health condition: Obesity, Lower Back Pain, and Cardiovascular Disease. These datasets, sourced

Table 3.1: Overview of Used Libraries (all web links accessed from January to April 2024).

Lib	Website	Description.
sklearn	https://scikit-learn.org	Predictive data analysis
pandas	https://pandas.pydata.org	Data manipulation
numpy	https://numpy.org	Scientific computing
matplotlib	https://matplotlib.org	Visualization
seaborn	https://seaborn.pydata.org	Data visualization
TensorFlow	https://www.tensorflow.org	ML platform
Keras	https://keras.io/api/backend/	Deep learning API
Imblearn	https://imbalanced-learn.org	Imbalanced data

from Kaggle.com, are well-regarded in healthcare research for their comprehensive data on patient demographics, health indicators, and clinical outcomes.

3.1.1 Obesity Dataset

Encompasses variables that shed light on dietary habits, physical activity levels, and genetic predispositions affecting obesity. This dataset is instrumental in understanding the multifaceted nature of obesity. It incorporates data on individuals' eating habits and physical conditions. With 2,111 entries and 17 features—spanning both numerical and categorical types—it offers a broad perspective on obesity-contributing factors. The dataset categorizes individuals into various obesity levels, from Insufficient Weight to Obesity Type III, making it suitable for multi-class classification tasks. Its size, larger than the Lower Back Pain dataset, presents a unique advantage for enhancing classification task performance. Furthermore, it allows for the examination of synthetic data generation techniques across a spectrum of feature types and more intricate classification challenges. For more information and data access, visit Kaggle: Obesity Prediction Dataset Link

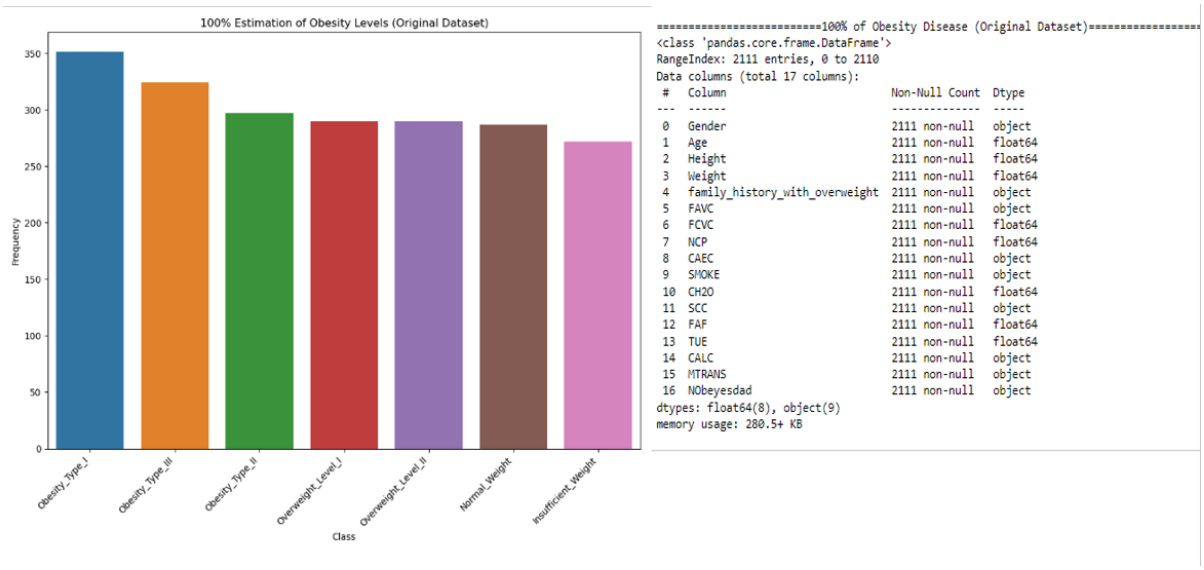
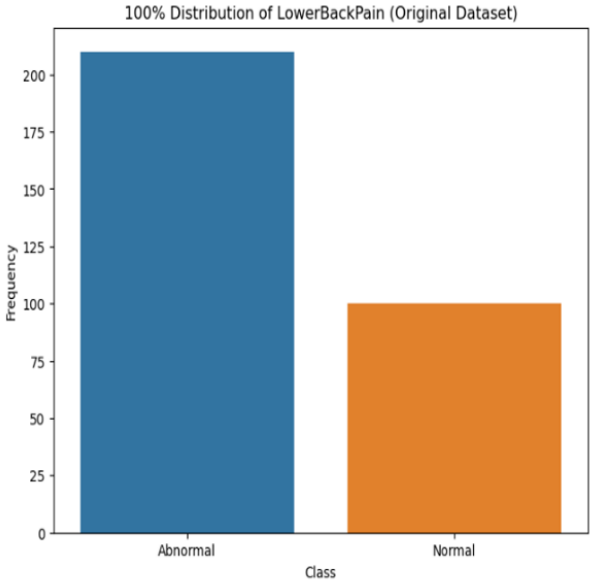


Figure 3.1: Obesity Prediction Dataset and Structure

3.1.2 Lower Back Pain Dataset

Features biomechanical attributes derived from orthopedic measurements. It is pivotal in identifying biomechanical factors contributing to lower back pain. This dataset is notable for its compact size, encompassing just 310 instances across 13 variables. It presents a unique blend of features, predominantly numerical, with a singular binary class label aimed at facilitating predictive analyses. This collection is meticulously curated to encapsulate a variety of factors implicated in lower back pain or lumbago, spanning a broad spectrum from muscular and ligament strains to nerve compression and skeletal irregularities within the lumbar region. It aims to aid in the identification of unusual biomechanical patterns that could potentially signal the onset of lumbago. Given its concise dataset size, it emerges as an exemplary model for testing the efficacy of synthetic data augmentation in enhancing analytical performance. Moreover, its class imbalance mirrors a prevalent issue within healthcare datasets, thereby reinforcing its relevance to this research's objectives. For access to the dataset, visit Kaggle: Lower Back Pain Symptoms Dataset Link.



==========100% of LowerBackPain (Original Dataset)==========

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 310 entries, 0 to 309
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   pelvic_incidence        310 non-null    float64
 1   pelvic_tilt             310 non-null    float64
 2   lumbar_lordosis_angle   310 non-null    float64
 3   sacral_slope            310 non-null    float64
 4   pelvic_radius           310 non-null    float64
 5   degree_spondylolisthesis 310 non-null   float64
 6   pelvic_slope            310 non-null    float64
 7   direct_tilt             310 non-null    float64
 8   thoracic_slope          310 non-null    float64
 9   cervical_tilt           310 non-null    float64
 10  sacrum_angle            310 non-null    float64
 11  scoliosis_slope         310 non-null    float64
 12  Class_att               310 non-null    object
dtypes: float64(12), object(1)
memory usage: 31.6+ KB
```

Figure 3.2: Lower Back Pain Symptoms Dataset and Structure

Table 3.2: Summary of Datasets Used in the Study

Dataset Name	Total Records	Numerical Features	Categorical Features	Total Features
Lower Back Pain Symptoms	310	12	1	13
Estimation of Obesity Levels	2,111	8	9	17
Cardiovascular Disease	70,000	5	6	11

3.1.3 Cardiovascular Disease Dataset

Comprises clinical parameters, such as blood pressure and cholesterol levels, alongside lifestyle factors, offering insights into the prevalence and predictors of cardiovascular diseases. Available on Kaggle, this dataset stands out due to its extensive volume, featuring 70,000 entries adorned with both objective and subjective data across 11 variables, blending medical findings with patient-reported information. With a mix of 5 numerical and 6 categorical attributes, it lays a robust foundation for predictive analyses concerning cardiovascular diseases. Notably, it assumes a pivotal role in our study by acting as a comparative benchmark, enabling an assessment of the efficacy of synthetic data generation across varying dataset

magnitudes. This comparative approach underscores the dataset's utility in validating the performance of synthetic data techniques not only within large-scale data environments but also when applied to datasets of a more modest scope. This ensures a well-rounded examination of the synthetic data generation's capabilities. For access to the dataset, visit Kaggle: Cardiovascular Disease Dataset Link

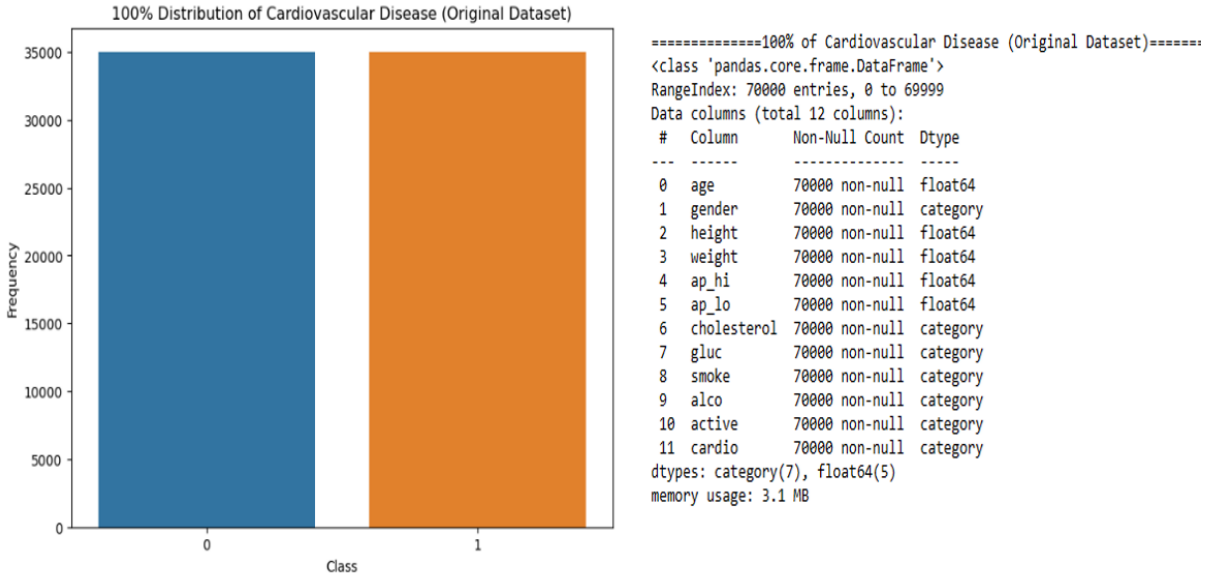


Figure 3.3: Cardiovascular Disease Dataset and Structure

This chapter meticulously details our approach from the initial data handling to the complex processes of data synthesis and evaluation. For hands-on examples, scripts for data preprocessing, and model configurations, explore our: GitHub Repository.

3.2 Data Preprocessing

This study utilized a comprehensive dataset segmented into two primary components respectively: original data and control data. Prior to synthetic data generation, data preprocessing steps were meticulously applied to each dataset, ensuring data cleanliness, normalization, and partitioning into the said two main component data. The original dataset, comprising 80% of the total data, served as the foundation for generating synthetic data. In contrast, the remaining 20%, referred to as control data, was reserved for evaluating both the privacy risks associated with the synthetic dataset and the utility as well.

3.2.1 Partitioned 80% Original Datasets

The partitioned 80% of original datasets form the cornerstone of our synthetic data generation process. These datasets are meticulously chosen subsets of the larger original datasets as accounted in the above section, serving as the basis for creating synthetic versions that mirror the real-world complexities and nuances of the data. The selection of these datasets follows a principled approach aimed at preserving the integrity and distributional characteristics of the original data, ensuring that the synthetic datasets generated thereafter retain their utility for research and analysis purposes.

Purpose and Utilization:

- Foundation for Synthetic Data: This partition acts as the direct input for synthetic data generation tools, ensuring that the resultant synthetic datasets are deeply rooted in the empirical reality captured by the original data.

- Benchmark for Synthetic Data Evaluation: Beyond serving as a basis for generation, these datasets enable a critical comparison between synthetic and original data. This comparison is vital for assessing the fidelity of synthetic datasets in terms of statistical properties and predictive capabilities.

Creating the Partition:

- Data Splitting Strategy: To prepare for synthetic data generation, the original dataset undergoes a partitioning process where 80% is earmarked for the generation phase. This strategic split is designed to balance the need for comprehensive synthetic data generation while reserving a substantial portion for control and validation purposes.

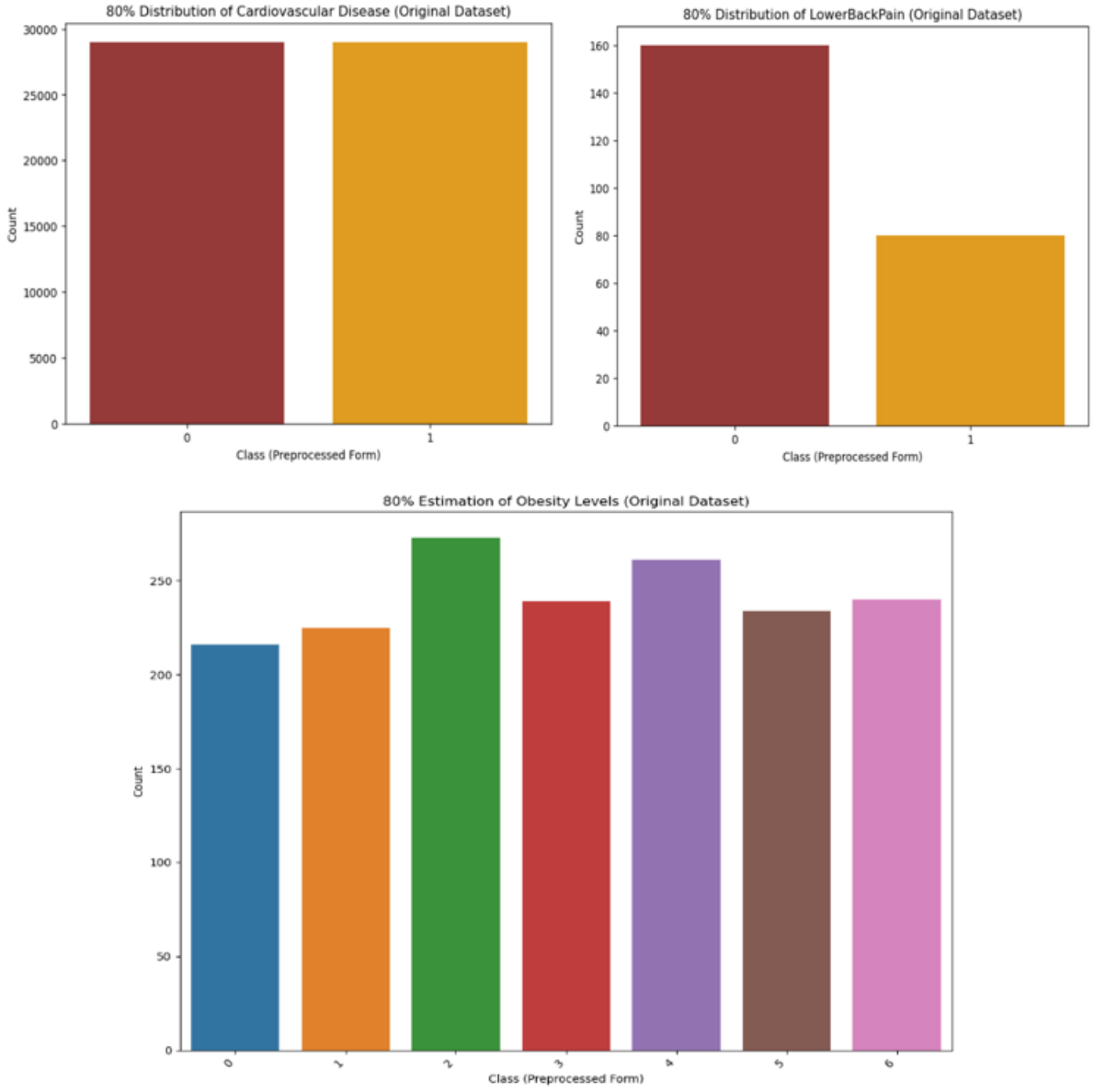


Figure 3.4: 80 Percent of the Variant Used Original Datasets

3.2.2 Partitioned 20% Control Datasets

The control dataset, constituting the remaining 20% of the original data, is pivotal in evaluating the privacy preservation capabilities of synthetic datasets. Its role extends beyond a mere

passive benchmark; it actively participates in the privacy risk assessment by serving as a litmus test for identifying potential privacy breaches.

Significance and Application:

- Measuring Privacy Leakage: By contrasting attack success rates on this dataset with those on the synthetic and original datasets, we discern the nature of information leakage, distinguishing between utility-driven insights and genuine privacy vulnerabilities.
- Authenticating Evaluation Rigor: The control dataset ensures that our assessment methodologies are grounded, offering a fair and unbiased evaluation of the privacy-preserving properties of synthetic data.

Compilation of the Control Dataset:

- Ensuring Exclusivity: The control dataset is compiled to ensure that it consists exclusively of data points not utilized in the synthetic dataset's creation. This exclusivity is crucial for maintaining the integrity of the privacy risk assessment process.
- Reflective of Original Data: Despite its separation from the generation process, the control dataset is representative of the overall data distribution, ensuring that conclusions drawn from its analysis are applicable to the broader dataset.

3.3 Benchmark for Comparison: The Original Dataset

In this study, we adopt a systematic approach to dataset selection and preparation, which is foundational to our research's integrity and validity. Among the various datasets utilized, we designate the original datasets as our primary benchmarks for comparison against the generated synthetic datasets. This section elucidates the rationale behind this choice and outlines the significance of a benchmark in data analysis. The original dataset, from which synthetic datasets are generated, serves as the cornerstone for our comparative analysis. This benchmarking strategy is predicated on several key considerations:

1. **Accuracy and Authenticity:** The original dataset encapsulates real-world complexities, embodying the true distributions and inter-variable relationships. As such, it offers an unaltered reflection of reality against which the synthetic data's fidelity can be assessed. This accuracy and authenticity make it an indispensable reference point for evaluating the preservation of statistical properties and utility in synthetic datasets.
2. **Utility Evaluation:** Central to our synthetic data generation endeavor is the retention of the original dataset's utility while safeguarding privacy. The original dataset, therefore, acts as a gold standard for evaluating how well synthetic datasets mimic its statistical properties and maintain analytical utility without compromising data subjects' privacy.
3. **Privacy Risk Assessment:** Employing the original dataset as a benchmark facilitates a rigorous assessment of the anonymization techniques applied to generate synthetic data. This comparison is crucial for determining the synthetic datasets' efficacy in obfuscating individual identities and mitigating the risk of sensitive information inference.
4. **Standardization of Evaluation:** By benchmarking against the original dataset, we introduce a standardized framework for evaluating synthetic data generation techniques. This standardization is vital for a systematic and objective comparison across different methodologies, enabling a clear articulation of progress and challenges in synthetic data research.

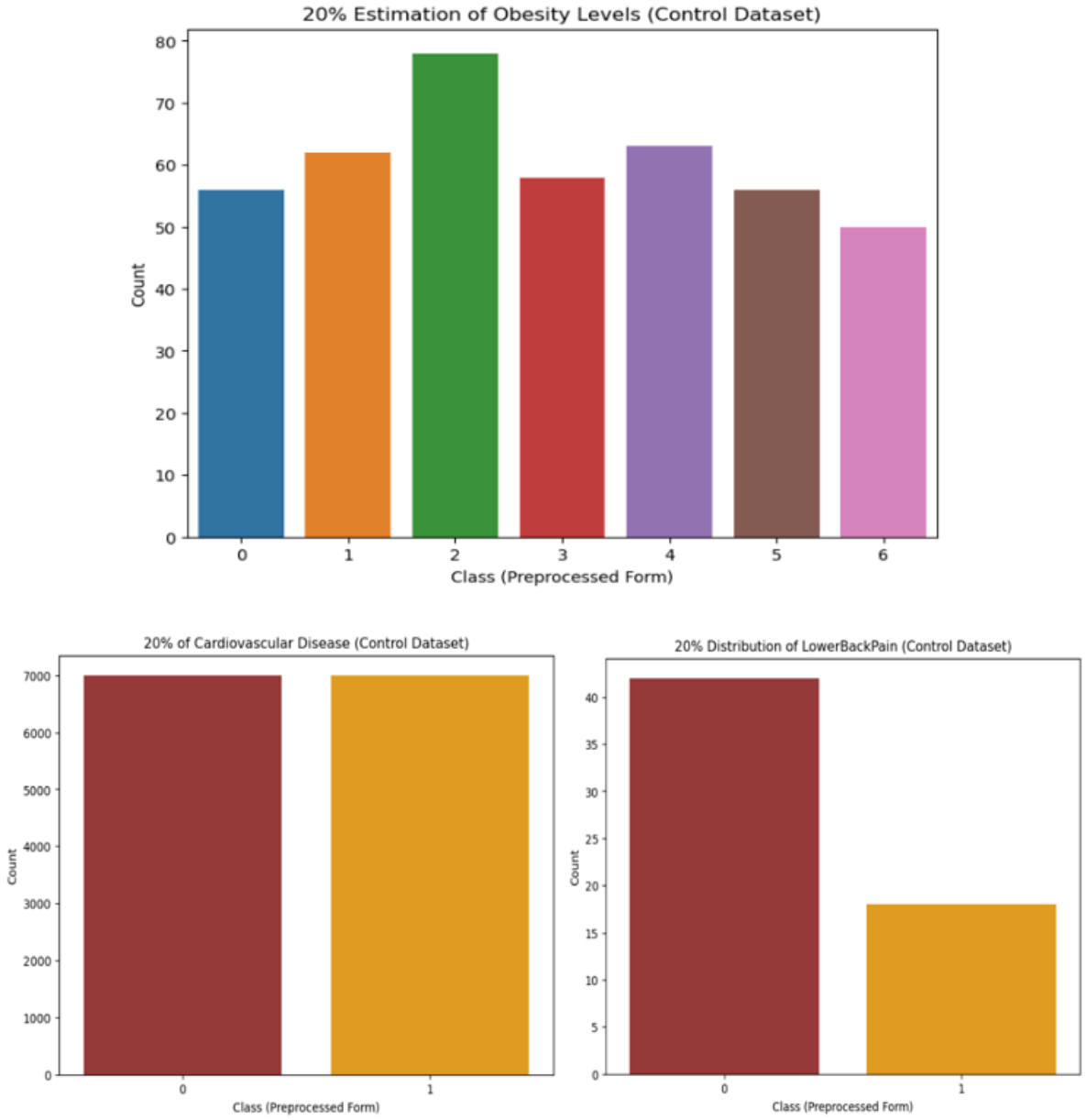


Figure 3.5: 20 Percent of the Variant Used Control Datasets

Establishing a benchmark for comparison is imperative for the methodological rigor of our study. The original dataset's role as this benchmark underpins our evaluative processes, ensuring our analyses are grounded in accuracy, utility, and privacy considerations. Through this benchmarking approach, we aim to contribute substantively to the discourse on synthetic data generation, offering insights that balance the dual imperatives of data utility and privacy.

3.4 Model Selection

In this project, we aimed to generate synthetic datasets that closely mirror the statistical properties of the original data while ensuring the preservation of individual privacy. To achieve this, we selected Autoencoders (AE) and Variational Autoencoders (VAE) as our primary models for synthetic data generation. These models were chosen due to their proven capability in learning deep representations of data, which is crucial for generating high-quality synthetic instances.

Autoencoder (AE) and Variational Autoencoder (VAE) models serve distinct purposes within our methodology. AE's straightforward encoding-decoding architecture makes it suitable for capturing and replicating the complex distributions of our datasets. In contrast, VAE introduces a probabilistic approach to encoding, offering the advantage of generating more diverse data samples, which is vital for exploring the utility and privacy dimensions of synthetic datasets.

3.4.1 Evaluation Framework Overview

Our methodology employs a multi-dimensional evaluation framework to rigorously assess the generated synthetic datasets across resemblance, classification performance, and privacy preservation. Inspired by established metrics, our framework leverages Python libraries such as NumPy for numerical operations, Pandas for data manipulation, Scikit-learn for machine learning models, and Seaborn for data visualization to ensure a thorough analysis.

Resemblance to Original Data

To ascertain the synthetic data's fidelity to real-world complexities, we conducted a series of statistical checks:

- **Basic Statistical Checks:** We began by quantitatively comparing column-wise means and standard deviations between the real and synthetic datasets, using visual plots to ensure the synthetic data adheres to the same univariate statistical characteristics as the real data. These statistical comparisons were done using metrics such as the KS-Test, P-Value, F-test, and T-test, which are mentioned in Chapter 2 of this report. The KS-Test and P-Value metrics help assess the distributional similarity of each feature between the original and synthetic datasets. The F-test and T-test compared variances and means, respectively, to identify significant differences between the original and synthetic datasets.
- **Correlation Analysis:** Employing the Pearson correlation coefficient and correlation heatmap, we measured how closely the synthetic data captures the intricate inter-variable relationships inherent in the real data. A near-zero difference matrix between the real and synthetic datasets' correlations indicated successful mimicry of these relationships. Metrics are mentioned in Chapter 2 of this report.
- **Feature Distribution Comparison:** Through overlaying feature distributions and cumulative sums of the real and synthetic data, we visually confirmed the synthetic data's accurate representation of the original data's statistical properties. Discrepancy metrics like KSComplement and TVComplement further quantified these similarities.

Classifier Performance Evaluation

In the realm of machine learning predictions, the synthetic data's utility was benchmarked against its real counterpart. Utilizing classifiers such as RandomForest, Gradient-Boosting, Decision Trees, Logistic Regression, MLP (Multilayer Perceptron), K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC) and XGBoost, we explored the synthetic data's ability to sustain the original dataset's predictive power across various tasks.

- **Training and Testing Phases:** Our approach, encompassing Training on Synthetic, Testing on Real (TSTR), and its reverse, TRTR, provided a comprehensive view of synthetic data performance. This strategy ensured a detailed comparison across real, synthetic, and augmented datasets.
- **Evaluation Metrics:** Cross-validation techniques and a combination of F1 scores and ROC curves offered a nuanced understanding of classifier performances. These metrics,

particularly adept at navigating the complexities of healthcare data, facilitated a balanced evaluation of precision, recall, and the trade-off between sensitivity and specificity.

Through this meticulously crafted methodology, we aim to generate synthetic datasets that not only closely resemble their original counterparts in statistical properties but also uphold the highest standards of privacy, ensuring their utility in sensitive applications like healthcare without compromising individual data subjects' privacy.

Finally, we evaluated the **privacy risk** of the generated synthetic data using **Anonymeter evaluators**. These evaluators provided a structured approach to assess risks related to singling out, linkability, and inference, aligning with GDPR's guidelines for privacy risk assessment in synthetic data. The choice of Anonymeter as our evaluation framework was motivated by its comprehensive risk assessment capabilities, offering insights into potential privacy vulnerabilities inherent in the synthetic datasets.

Rationale for Model Selection: The selection of AE and VAE models was informed by their ability to generate synthetic data that is both useful for analytical purposes and resilient against privacy attacks. The variety of classifiers and statistical methods employed were chosen to ensure a multidimensional evaluation of the synthetic data, considering aspects such as accuracy, diversity, and statistical fidelity. This holistic approach ensures that the synthetic datasets produced not only serve as viable substitutes for the original data in terms of utility but also adhere to stringent privacy standards.

In summary, the methodology adopted in this project is underpinned by a strategic model selection process, designed to balance the dual objectives of data utility and privacy preservation. By meticulously choosing and evaluating our models and methodologies, we ensure that the synthetic datasets generated are not only statistically representative of the original data but also uphold the highest standards of individual privacy.

3.5 Implementing Data Generation Processes Leveraging AE and VAE

In the pursuit of generating high-quality synthetic datasets that mirror the complexity and distribution of real-world healthcare data, our project employed two sophisticated neural network architectures: Autoencoders (AE) and Variational Autoencoders (VAE). This is a continuation of the foundational background documented in Chapter 2. These models were pivotal in learning deep representations of the data, facilitating the generation of new synthetic instances (data) that preserve the statistical properties of the original datasets and structure of the original data while ensuring the privacy of individual records. These sections will delve into the implementation specifics, including the choice of architecture, training details, and the criteria for evaluating the synthetic data's fidelity to the original datasets.

3.5.1 Autoencoders Configuration (AEC)

In synthetic data generation, AE was employed to capture and model the complex distributions of the healthcare datasets, as discussed in the preceding chapter.

- **Network Architecture Choice:** The AE architecture was carefully chosen based on the specific features of each dataset, considering the dimensionality and diversity of the data. It consisted of an encoder network that compressed input data into a lower-dimensional latent space and a decoder network that reconstructed the data from this latent representation. The encoder was structured with dense layers, transitioning from an input dimension to a bottleneck layer with a reduced dimensionality, thus capturing the essential features. The decoder mirrored this architecture in reverse, aiming to reconstruct the input data as closely as possible.

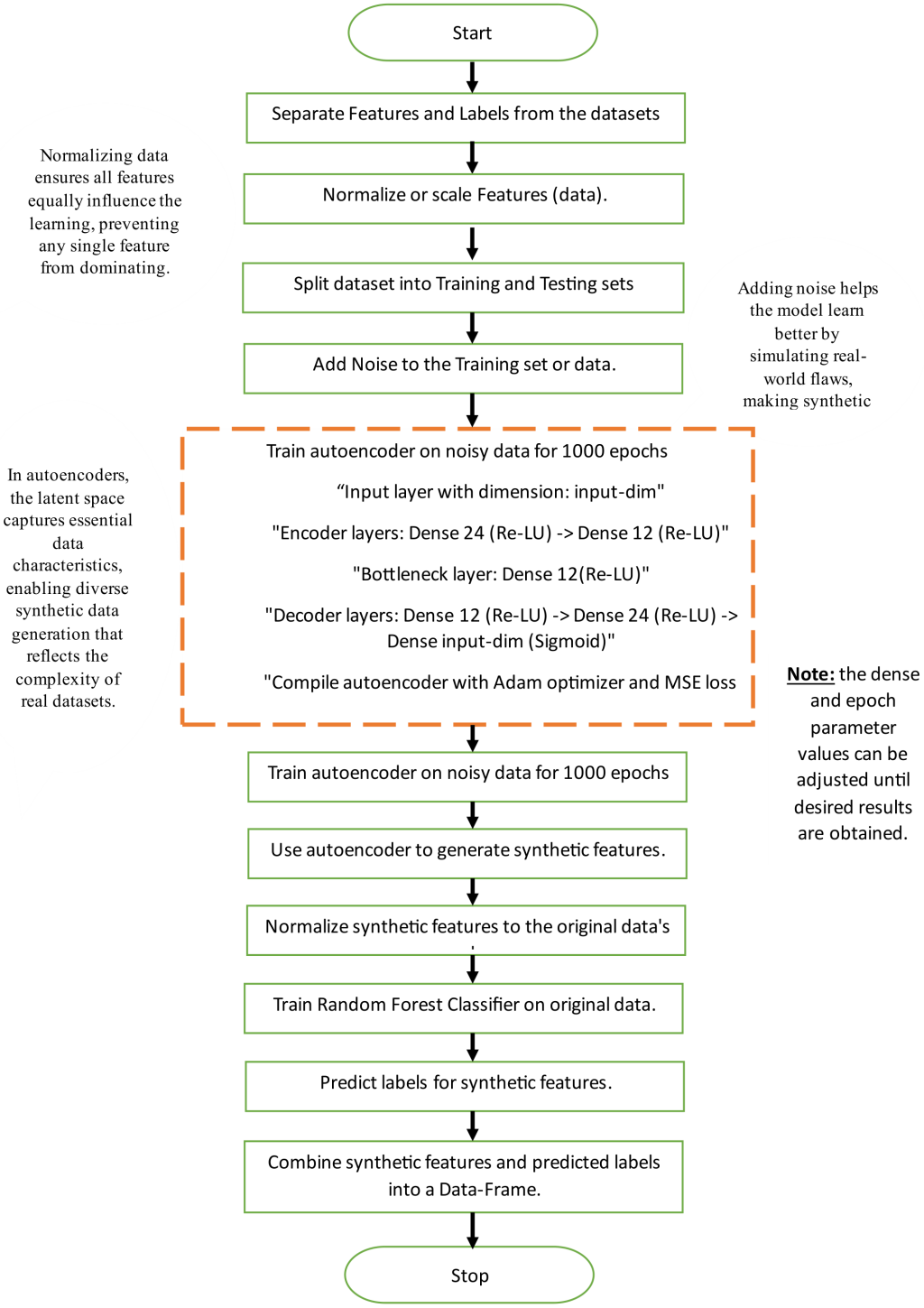


Figure 3.6: Autoencoder Flowchart for Synthetic Data Generation

- **Training Process Details:** The AEs were trained using a subset of the original datasets, excluding the control set. The training involved minimizing the reconstruction error, defined as the mean squared error (MSE) between the original data and its

reconstruction. The AE was trained using the Adam optimizer, with a learning rate of 0.001, over 1000 epochs, showcasing the model's ability to learn intricate data patterns over iterative training cycles. Sigmoid was used as the activation function.

- **Parameter Settings:** The AE utilized a latent dimension size of 12 and 16 for the bottleneck layer, for the respective datasets used, selected to balance between data compression and reconstruction quality. The model's capacity was adjusted to prevent overfitting, ensuring that the generated synthetic data was both diverse and representative of the original dataset.
- **Fidelity Evaluation:** To assess the fidelity of the synthetic datasets, we compared distributions of key variables, correlation matrices, and other statistical metrics between the synthetic and original datasets. This comprehensive evaluation ensured the utility of the synthetic data for subsequent analyses and model training. This will be documented in Chapter 4 below.

Following dataset organization and the splitting of the respective 80% data (cardiovascular, obesity, and lower-back pain datasets) into training and testing sets, we ensured a rigorous model performance evaluation. Implementing normalization via MinMaxScaler was critical in standardizing feature scales across the datasets, thereby streamlining the model training process. A pivotal enhancement in our approach was integrating noise into the training dataset. By introducing a noise factor of 0.05, we aimed to amplify the autoencoder's generalization capabilities, thus elevating the synthetic dataset's quality and utility.

The autoencoder's architecture was meticulously designed, comprising an encoder that condensed input features into a dense latent representation of 32 units through a bottleneck layer, utilizing ReLU activation functions for both the encoding and bottleneck stages. This encoder was followed by a decoder aiming to reconstruct the input from its compressed form. Specifically, the encoder network featured two dense layers of 24 and 12 units respectively, each employing ReLU activation functions. Post-bottleneck, the decoder network mirrored this structure in reverse, starting from 12 units, escalating to 24 units, and culminating in reconstructing the original input dimension with a sigmoid activation function in the final layer.

Training the autoencoder on the noise-enhanced dataset was conducted meticulously, focusing on optimizing the model's proficiency in synthesizing data. Upon the culmination of training, the autoencoder was utilized to generate synthetic features from the scaled original dataset, a process pivotal in ensuring that the synthetic data precisely mirrored the original data's statistical characteristics.

A RandomForestClassifier, previously calibrated on the original dataset, was employed to infer labels for the synthetic features, facilitating the assembly of an exhaustive synthetic dataset poised for subsequent analyses. This synthetic dataset, composed of features and labels predicted by the classifier, represents a significant stride towards our aim of synthesizing data that retains the intrinsic properties of the original dataset while safeguarding privacy.

3.5.2 Variational Autoencoders Configuration (VAEC)

This approach not only facilitates data generation but also imbues the synthetic data with variability, enhancing privacy.

- **Network Architecture Choice:** The VAE architecture was similarly tailored to the intricacies of the healthcare datasets. The VAE introduced a probabilistic twist to the conventional AE architecture. Besides encoding an input into a latent representation, the encoder of a VAE maps inputs to a distribution in latent space. The decoder then samples from this distribution to generate synthetic data, enhancing diversity.

Algorithm 1: Autoencoder-Based Synthetic Data Generation

Input: Original healthcare dataset (features, labels)
Output: Synthetic healthcare dataset (synthetic_features, synthetic_labels_predicted)
(features_scaled, labels) ← Preprocess(features, labels)
$(X\_train\_scaled, X\_test\_scaled, y\_train, y\_test)$ ← DataSplit(features_scaled, labels)
X_train_noisy ← AddNoise(X_train_scaled)
autoencoder ← TrainAE(X_train_noisy, X_train_scaled)
synthetic_features_scaled ← GenerateSynthetic(autoencoder, features_scaled)
synthetic_features ← PostProcess(synthetic_features_scaled)
classifier ← TrainClassifier(X_train, y_train)
synthetic_labels_predicted ← PredictLabels(classifier, synthetic_features)
synthetic_data_df ← AssembleDataset(synthetic_features, synthetic_labels_predicted)
return synthetic_data_df

- **Training Process Details:** The VAE's loss function comprised two terms: the reconstruction loss (similar to AE) and the Kullback-Leibler divergence, promoting efficient encoding in latent space. The training process, conducted over 15000 epochs, leveraged the Adam optimizer with a learning rate of 0.001, and Sigmoid as the activation function. This ensured that while the synthetic data retained similarity to the original, it also introduced variability to prevent direct re-identification.

- **Parameter Settings:** The VAE's latent dimension was set to 12 and 16, respectively, depending on the dataset used, optimizing the model's ability to generate a wide array of synthetic instances. This dimensionality was chosen after extensive testing to ensure an optimal trade-off between model complexity and synthetic data quality.

- **Fidelity Evaluation:** Beyond statistical similarity, the evaluation of VAE-generated data also considered the diversity and novelty of the synthetic instances. Metrics such as sample uniqueness and entropy were utilized alongside traditional statistical comparisons to validate the synthetic data's quality. This will be documented in Chapter 4 below.

In creatinging synthetic datasets with a Variational Autoencoder (VAE) model, we began by preprocessing features and targets from the respective 80% (cardiovascular, obesity, and lower-backpain datasets). This step was crucial for setting the stage for later analyses.

The VAE model, distinguished by its probabilistic encoding, was chosen for its ability to produce diverse synthetic data. We normalized features using MinMaxScaler to standardize data across the datasets and introduced noise to enhance the model's generalization, aiming to improve the synthetic data's realism. The VAE's architecture, featuring an encoder-decoder network and a 12-dimensional latent space, was central to its performance. The encoder compressed data into a latent representation, while a sampling function in the encoder introduced variability, enabling the generation of diverse synthetic samples. (Note that the dimensional latent spaces used are 12 and 16 dimensions depending on the dataset used). The obesity dataset is composed of 17 features). After training, the VAE synthesized data that closely resembled the original datasets' statistical properties. This included generating synthetic data samples conditioned on class labels, ensuring the resultant dataset reflected real-world data's complexity and variability.

The synthetic data was then rescaled to match the original scale, ensuring consistency for analyses. This streamlined approach from preprocessing to synthetic data generation exemplifies our commitment to balancing data privacy with utility, producing a synthetic dataset that addresses privacy concerns while maintaining relevance for research.

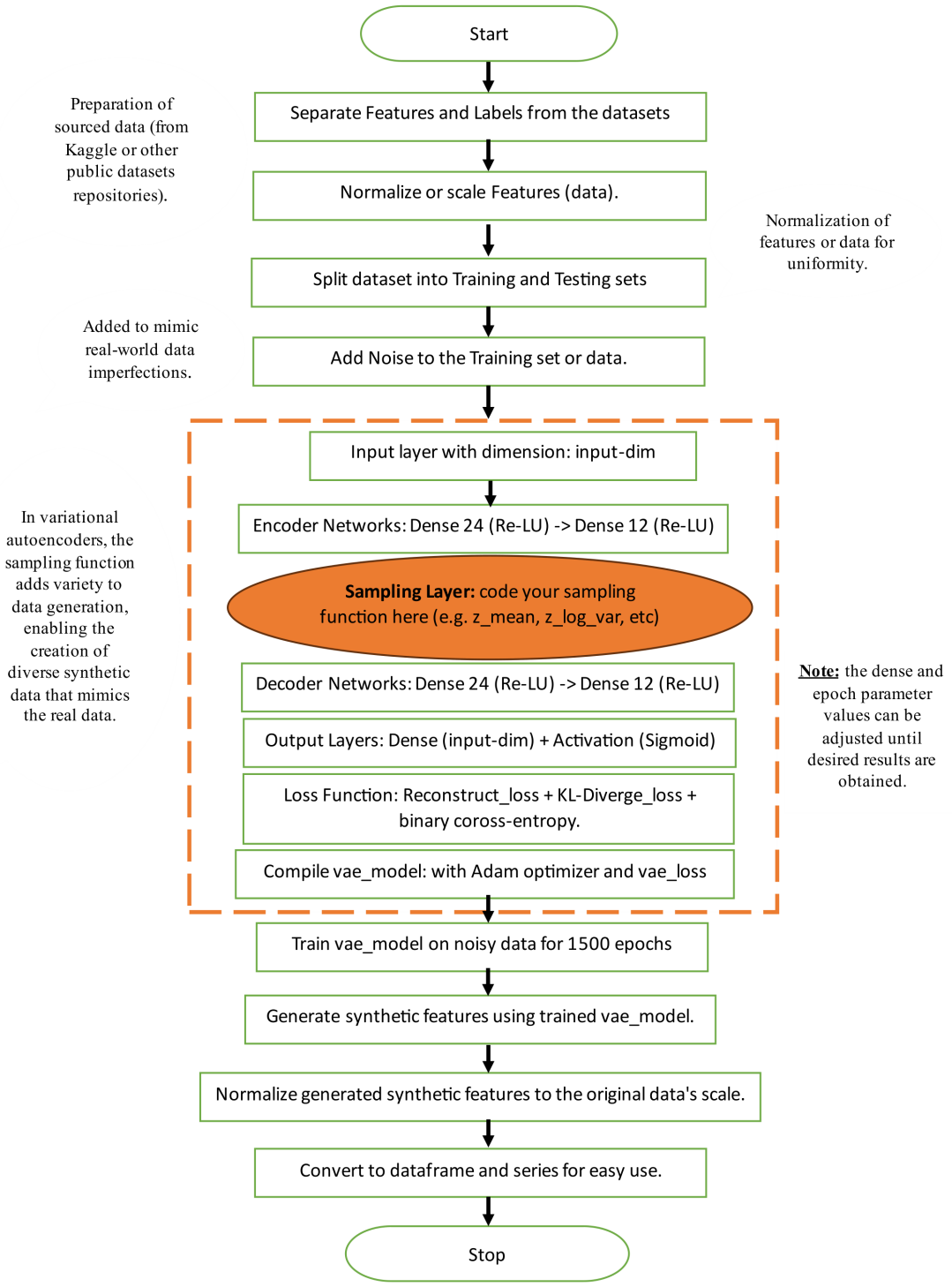


Figure 3.7: Variational Autoencoder Flowchart for Synthetic Data Generation

The operational processes detailed above Figure3.6 and Figure 3.7, illustrate the meticulous steps undertaken to generate synthetic datasets using AE and VAE. This approach not only addresses the privacy concerns inherent in utilizing real-world data but also ensures the utility of the generated synthetic datasets for further research and application. The subsequent

sections will delve into the privacy risk assessment, model performance evaluation, and comparison with existing literature, providing a comprehensive analysis of the efficacy of synthetic data in preserving privacy while maintaining utility.

Algorithm 2: Variational Autoencoder-Based Synthetic Data Generation

Input: Original dataset with features and labels
Output: Synthetic dataset with generated features and labels
Preprocess the dataset:
    a. Split the dataset into training and testing sets.
    b. Normalize features using MinMaxScaler.
Add noise to the training data to create noisy inputs.
Construct the VAE architecture:
    a. Define the input layer for features and labels.
    b. Embed labels and concatenate with input features.
    c. Build the encoder network to learn latent representations:
        - Dense layers with ReLU activation.
        - Compute mean and log variance of latent space.
        - Apply the sampling function to generate latent vectors.
    d. Build the decoder network to reconstruct input features:
        - Dense layers with ReLU activation and sigmoid output.
Define the VAE loss function:
    a. Reconstruction loss: Binary cross-entropy between input and output.
    b. KL divergence loss: Measures divergence from the prior distribution.
Compile and train the VAE model on the noisy data.
Generate synthetic data:
    a. Sample latent vectors from the normal distribution.
    b. Use the decoder to generate synthetic features from sampled latent vectors.
Rescale synthetic data back to original feature space.
Convert synthetic features and labels to DataFrame and Series for use.
(Optional) Train a classifier on the original data and predict labels for the synthetic data.

3.6 Privacy Risk Assessment with Anonymeter Framework

Utilizing the Anonymeter tool, the privacy risks associated with the synthetic datasets were evaluated across three dimensions: Singling-Out, Linkability, and Inference Risks. This section will elaborate on the methodologies and algorithms employed in these assessments, continuing from the brief theoretical frameworks discussed in Chapter 2.

In the era of data-driven decision-making, the use of synthetic datasets has emerged as a powerful tool for preserving privacy while leveraging big data for analytics, machine learning, and research. Synthetic data generation involves creating artificial data that is statistically like real-world data but does not directly correspond to any real individual's information. This approach holds the promise of mitigating privacy risks, enabling organizations to comply with stringent data protection regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) without stifling innovation.

However, the generation of synthetic data introduces new challenges in ensuring that these datasets do not inadvertently reveal sensitive information about individuals. To address these challenges, tools, and frameworks for assessing privacy risks associated with synthetic datasets are critically needed. The Anonymeter framework represents a pivotal advancement in this domain, offering a comprehensive suite of evaluations designed to measure the privacy

preservation capabilities of synthetic datasets.

3.6.1 Simplified Overview: Overview of Anonymeter's Privacy Assessment Process

Introduction

In our quest to ensure the privacy of individuals within synthetic datasets, we leverage a powerful tool known as Anonymeter. Developed by Anonos, experts in data privacy, this tool offers an essential check against potential privacy risks that might arise from data anonymization processes [1, 10, 22, 32].

Simplified Explanation

At its core, Anonymeter undertakes a series of steps to ascertain how well our synthetic datasets protect individual privacy. Think of it as a guardian that meticulously examines the data, ensuring no individual can be identified from the information provided. It performs this by simulating potential attacks, akin to testing the locks on a door, to see if an intruder could gain access.

1. **Attack Phase:** Initially, Anonymeter simulates attacks on the synthetic dataset, akin (similar) to a friendly hacker testing our defenses.
2. **Evaluation Phase:** It then assesses how these simulated attacks fare when attempting to match synthetic data with real individuals' records.
3. **Risk Estimation Phase:** Finally, Anonymeter measures the effectiveness of these attacks on a separate dataset known as the control dataset to determine if any real risk of identification exists. Note: the control dataset is already discussed in the section above.

This process ensures that our synthetic datasets can be used safely for research without compromising individual privacy.

3.6.2 Detailed Technical Explanation: Technical Details of Anonymeter's Operational Process

Introduction to Technical Details

Delving deeper into the operational mechanics of Anonymeter, we uncover a sophisticated framework designed to evaluate and mitigate privacy risks in synthetic datasets. This section aims to provide a comprehensive understanding of the methodologies and algorithms that underpin these evaluations [1, 10, 22, 32].

Detailed Operational Process

Anonymeter's assessment process is a rigorous examination of how synthetic datasets stand against privacy breaches, detailed as follows:

1. **Attack Phase:** This phase involves executing sophisticated privacy attacks on the synthetic dataset. The tool employs algorithms to generate educated guesses or attacks that mimic potential malicious attempts to re-identify individuals. This process evaluates the dataset's vulnerability to singling out, linkability, and inference attacks.
2. **Evaluation Phase:** The subsequent phase compares the effectiveness of these attacks against the original dataset, offering a measure of the synthetic data's resemblance to real-world data and its susceptibility to privacy breaches. This comparison is pivotal in understanding the balance between data utility and privacy.

3. **Risk Estimation Phase:** Anonymeter extends its analysis to a control dataset, not involved in generating the synthetic data. This comparison helps isolate the insights gleaned purely from the dataset's utility from those indicating privacy leaks. It's a critical step in distinguishing between genuine privacy risks and benign data similarities.

The Triad of Attacks

Three distinct types of attacks underscore Anonymeter's methodology. In Anonymeter's evaluation, the **Main Privacy Attack** thinks of synthetic data as a tool for making smart guesses about real data. The goal here is to see if any real information can be picked out from the fake data, which in this case the synthetic data. Concurrently, the **Control Privacy Attack**, this time, uses the synthetic data to guess info about a separate, untouched set of data. It helps determine what's a helpful guess and what might be a privacy no-no. Complementing these, the **Baseline Attack**, at this point, we pretend to be a guesser who doesn't use the synthetic data at all, just making random guesses. It's like checking to see if just flipping a coin would be as good as using the synthetic data. This simply employs random guesses to establish a baseline for success rates, ensuring that the effectiveness of the main attack significantly surpasses that of uninformed guessing. This triad of attacks meticulously gauges the balance between data utility and privacy risk. [22, 32].

Aftermath of the Attacks: The heart of checking how safe Anonymeter's data is involves looking at how well these attacks work. It's all about seeing if the guesses are right and figuring out the risk of someone's private info being guessed. We pay special attention to how the smart guesses compare to the random guesses, which helps us see if the synthetic data gives away too much real info. This way, we make sure the synthetic data is useful but doesn't spill the beans on anyone's private details.

Evaluation Process

The core of anonymeter's evaluation lies in its focus on the three main privacy risks. Each risk is assessed through distinct evaluator classes:

- **Singling-Out Risk (Evaluator):** Tests the possibility of isolating a single individual's record within the dataset. The risk is that an individual can be singled out from a dataset, even if the data is supposed to be anonymized or synthetic. This analysis was conducted both in *univariate* and *multivariate* contexts to ascertain the risk of identifying individual data points within the synthetic dataset. A low risk here suggests that the synthetic dataset does not allow for easy re-identification of individuals based on a subset of attributes [22, 32].

- **Linkability Risk (Evaluator):** Assesses the potential risk of linking two or more distinct records, possibly across datasets to the same individual. This assessment determines the potential of linking individuals across the synthetic and original datasets, thus evaluating the risk of cross-dataset identification. Low linkability risk indicates that the synthetic dataset effectively disguises or alters connections that could be used to match records across datasets [22, 32].

- **Inference Risk (Evaluator):** The chance that sensitive information about individuals can be inferred, even when direct identifiers are removed. This dimension evaluates the likelihood of inferring sensitive attributes within the synthetic dataset, thereby assessing the risk of attribute disclosure. Low inference risk implies that the synthetic dataset protects against the derivation of sensitive information not intended to be disclosed [22, 32].

Algorithmic Operational Policies

For each privacy risk, anonymeter adopts a structured approach, leveraging algorithms to quantify risk levels effectively:

Algorithm 3: for Singling Out Evaluation

Result: Estimate the singling out risk.

```
Select a subset of records from the synthetic dataset;
foreach record in the subset do
    Attempt to match the record against the original dataset based on available attributes;
end
Calculate the success rate of matches to estimate the singling out risk;
```

Algorithm 4: for Linkability Evaluation

Result: Quantify the linkability risk.

```
Identify records in the synthetic dataset that share attributes with records in a separate
  control dataset;
Determine the proportion of these records that can be correctly linked back to the same
  individual in the original dataset;
Quantify the linkability risk based on the success rate of correct linkages;
```

Algorithm 5: for Inference Evaluation

Result: Assess the risk of sensitive information inference.

```
Utilize machine learning models to predict sensitive attributes of individuals based on
  other available data points in the synthetic dataset;
Compare these predictions against the actual values in the original dataset to assess the
  accuracy and potential for sensitive information inference;
The higher the predictive accuracy, the greater the risk of sensitive information inference;
```

Given the growing emphasis on data privacy and the increasing use of synthetic data for research, development, and analytics, tools like the Anonymeter are essential. They help balance the need for rich, usable data with the imperative to protect individual privacy. The methodologies employed align with the Anonymeter tool's usage policy, necessitating the preparation of three datasets with similar structures: the original dataset, the synthetic dataset, and the control dataset. The control dataset, derived from 20% of the original data, plays a pivotal role in the privacy risk assessment, offering a benchmark for evaluating the synthetic dataset's privacy preservation capabilities.

Summary

This chapter outlined and provided robust elaborative explanations of the methodological approach adopted in this study, encompassing dataset description, preparation, synthetic data generation, privacy risk assessment, and the evaluation of synthetic data utility and privacy. The methodologies were carefully chosen and implemented to address the project's core objectives, ensuring the generation of synthetic data that upholds privacy without compromising data utility.

Statistical Metrics and Their Uses

KS-Test (Kolmogorov-Smirnov Test)

The KS-Test compares two samples to determine if they come from the same distribution. It is non-parametric and works on continuous and discrete data. The test statistic quantifies the distance between the empirical distribution functions of the two samples. [16] [35]

F-test

The F-test assesses the equality of variances between two populations. It is often used in ANOVA (Analysis of Variance) to test the hypothesis that the means of several groups are equal, assuming the samples come from normally distributed populations with equal variances.

T-test

The T-test evaluates whether the means of two data samples are significantly different. It assumes that the samples are drawn from populations with identical variances (homoscedasticity) and follows a Student's t-distribution under the null hypothesis.

P-Value

The P-value quantifies the probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is correct. A small P-value (typically 0.05) indicates strong evidence against the null hypothesis, thus it is rejected.

MAE (Mean Absolute Error)

MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It's the mean of the absolute values of the differences between forecasted and observed values, offering a straightforward interpretation of prediction accuracy.

RMSE (Root Mean Squared Error)

RMSE is a quadratic scoring rule that measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation. RMSE is sensitive to outliers and gives a relatively high weight to large errors.

MSE (Mean Squared Error)

MSE measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

Mean

The mean is the average of a data set, calculated by adding all numbers in the data set and then dividing by the count of those numbers. It provides a central location for the data.

Variance

Variance measures how far a data set is spread out. It is the average of the squared differences from the Mean. A high variance indicates that the data points are spread out from the mean and from each other.

Std (Standard Deviation)

Standard Deviation is a measure of the amount of variation or dispersion in a set of values. A low standard deviation indicates that the values tend to be close to the mean, while a high standard deviation indicates that the values are spread out over a wider range.

Chi-Squared Test

The chi-square test is a statistical method used to determine whether there is a significant association between two categorical variables. It involves comparing the observed frequencies in each category of a contingency table with the expected frequencies calculated under the assumption that the variables are independent. If there are significant deviations between the observed and expected frequencies, the test suggests that the variables may be associated. This method was invented by the English statistician Karl Pearson in 1900, and it remains a fundamental tool in statistical analysis for testing independence and goodness of fit[61].

Chapter 4

Results

In this Chapter, we explore the utility and privacy preservation assessments of synthetic datasets generated through Autoencoders (AE) and Variational Autoencoders (VAE) across three distinct health conditions: obesity, cardiovascular disease, and lower back pain. The chapter is structured to provide detailed results for each condition, comparing the performance of AE-synthetic and VAE-synthetic datasets. Each dataset's utility is assessed in terms of its ability to mimic original data properties, while privacy preservation is evaluated through a series of assessments including univariate, multivariate, linkability, and inference risk evaluations. These assessments are visualized through bar and pie charts, which clearly depict success and failure rates for various security attacks against the synthetic data.

For a detailed explanation of the mathematical expressions and computational processes used in generating these pie charts, readers are referred to the appendix. The appendix section, particularly 6, elaborates on the methods used to calculate the values represented in the pie charts across all three datasets. This includes the step-by-step calculation of success and failure rates for the Main, Baseline, and Control attacks, providing a comprehensive understanding of how data security effectiveness is quantified in this research.

4.1 Overview of Result Activities

In refining our autoencoder (AE) and variational autoencoder (VAE) models to enhance their privacy preservation capabilities, we implemented several modifications. Initial assessments were conducted after using the Anonymeter tool to reveal specific privacy vulnerabilities, leading us to adjust the latent space dimensionality of our models. This change significantly obscured detailed input data features, reducing the risk of data reversibility.

Additionally, we employed dropout as a regularization technique and optimized the architecture by adjusting the number of units in Dense layers—transitioning from Dense(64) to Dense(32) or fewer. These modifications prevented the models from memorizing or reproducing identifiable details from the training data, thereby enhancing privacy.

We also perturbed the training dataset prior to training to mask key identifiers, further diminishing the resemblance between the synthetic and original datasets. The effectiveness of these adjustments was continuously evaluated with the Anonymeter tool, enabling us to monitor the impact on privacy preservation.

Quantitative Evaluation of Data Utility

For detailed information on the specific implementations and evaluations, please visit our GitHub Repository. Following the modifications, we conducted a comprehensive quantitative assessment comparing the statistical properties and machine learning model performances on the synthetic versus original datasets. The utility of the synthetic data for analytical tasks was

Please visit our GitHub for the complete codebase and documentation:
https://github.com/Kelechiede/Enhancing_Data_Security_with_AE_and_VAE_Synthetic_Data_Generation

evidenced through metrics such as accuracy, F1 score, and AUC-ROC.

Examples and Case Studies

The positive outcomes of our quantitative evaluations are further illustrated through case studies where the synthetic data was employed in predictive modeling tasks. These examples highlight the practical utility of the synthetic data and demonstrate how our refined models mitigate privacy risks effectively.

Privacy Risk Assessment

The efficacy of the synthetic datasets in preserving privacy was rigorously evaluated using the Anonymeter tool. The graphical results provided detailed insights into risks such as singling out individuals, linkability, and inference, crucial for identifying and addressing any remaining privacy vulnerabilities.

Trade-offs and Balancing Act

Our project navigated the complex trade-offs between maximizing data utility and upholding stringent privacy standards. This section explores the strategies we used to balance these competing objectives, ensuring the synthetic data remains useful for analytical purposes while complying with privacy regulations. Our approach highlights the delicate balance essential in synthetic data generation.

4.2 Data Utility of Obesity Data

4.2.1 Comparative Analysis of Original and AE Synthetic Obesity Datasets: A Multi-Faceted Evaluation Using AE Model

In our analysis, we've employed both Autoencoder (AE) and Variational Autoencoder (VAE) models to generate synthetic data, aiming to closely replicate the statistical properties of the original obesity dataset. This dataset encompasses a wide array of features, including demographic information, dietary habits, lifestyle factors, and obesity classification, which are crucial for understanding the multifaceted nature of obesity.

The results, particularly the P-values derived from the Kolmogorov-Smirnov (KS) test, serve as a quantitative measure to assess the distributional similarity between the original and synthetic datasets for each feature. A low P-value indicates a significant difference between the two distributions, suggesting areas where the synthetic data generation process may require refinement. Conversely, a high P-value suggests no significant difference, indicating that the synthetic data closely mirrors the original data in terms of distribution.

The choice of AE and VAE models was driven by their unique capabilities. AEs are known for their effectiveness in learning compressed representations of data, which is crucial for accurately capturing the underlying patterns in the obesity dataset. VAEs, on the other hand, introduce a probabilistic twist to the encoding process, allowing for the generation of new data points that adhere to the learned distribution. This probabilistic approach is particularly beneficial for ensuring the diversity and realism of the synthetic data.

Through this systematic presentation of results, we aim to provide a comprehensive understanding of the fidelity of the synthetic data to the original dataset.

The Kolmogorov-Smirnov (KS) test is a statistical method used to compare two datasets to determine if they have different distributions. It's particularly useful for checking how well a sample matches a reference probability distribution or for comparing two samples to see

if they come from the same distribution. The test calculates a statistic that measures the largest distance between the empirical cumulative distribution functions of the two datasets. A significant result suggests that the datasets have different distributions, while a non-significant result indicates similar distributions. The KS test is non-parametric, meaning it doesn't assume a specific type of distribution for the data, making it versatile for various types of data analysis.

The P-values presented in Table 4.1 were obtained from the Kolmogorov-Smirnov (KS) test, a non-parametric test used to determine if there is a significant difference between the distributions of two datasets. This test was specifically chosen for its ability to compare the empirical distribution functions of both the original and synthetic datasets without making any assumptions about the distribution of the data. The KS test is particularly useful in this context as it provides a straightforward metric (the P-value) to evaluate the similarity between the original data's distribution and that of the synthetic data generated by the Autoencoder. A low P-value indicates a significant difference between the distributions, suggesting that the synthetic data may not accurately capture the original data's statistical properties. Conversely, a high P-value supports the null hypothesis that there is no significant difference between the distributions, indicating a closer match between the synthetic and original datasets.

P-values are used to evaluate the likelihood of observing the test results under the null hypothesis that there is no significant difference between the distributions of the original and synthetic datasets for each feature tested. A low P-value suggests that the null hypothesis can be rejected, indicating a significant difference between the distributions of the original and synthetic data for that feature. Conversely, a high P-value suggests that there is not enough evidence to reject the null hypothesis, indicating that the distributions of the original and synthetic data are similar for that feature.

The larger P-values observed for Height and Weight, each above 0.05, suggest that the differences between the distributions of these features in the original and synthetic datasets are not statistically significant, indicating a good match in the synthetic representation. Conversely, the P-value for NObeyesdad, extremely high at 0.993, also supports the lack of significant differences between the original and synthetic datasets for this variable.

Table 4.1: Statistical Metrics and P-Value Comparison Between Original and AE Synthetic Obesity Dataset (Significance Highlighted)

Feature	**Orig Mean**	**AE Syn Mean**	**Orig Std Dev**	**AE Syn Std Dev**	**P-Value**
Gender	0.507	0.506	0.500	0.499	$5.95 \times 10^{-187***}$
Age	24.449	24.390	6.477	6.434	$8.50 \times 10^{-3*}$
Height	1.702	1.700	0.093	0.092	0.502
Weight	86.598	87.327	26.099	26.090	0.119
FHOW	0.819	0.820	0.385	0.384	$< 0.001^{***}$
FAVC	0.888	0.889	0.315	0.314	$1.43 \times 10^{-250***}$
FCVC	2.422	2.427	0.537	0.528	$1.25 \times 10^{-73***}$
NCP	2.686	2.640	0.783	0.779	$2.59 \times 10^{-101***}$
CAEC	1.855	1.845	0.478	0.479	$1.40 \times 10^{-162***}$
SMOKE	0.023	0.021	0.150	0.142	$< 0.001^{***}$
CH2O	2.006	2.007	0.610	0.598	$1.94 \times 10^{-9***}$
SCC	0.044	0.043	0.205	0.203	$< 0.001^{***}$
FAF	1.004	1.030	0.840	0.850	$1.36 \times 10^{-27***}$
TUE	0.644	0.634	0.603	0.597	$1.52 \times 10^{-55***}$
MTRANS	2.355	2.326	1.272	1.254	$3.71 \times 10^{-192***}$
NObeyesdad	3.046	3.034	1.958	1.936	0.993

Note: $^{*}P < 0.05$, $^{***}P < 0.001$ indicate significant differences.

Understanding Statistical Metrics and Computations

NOTE: The explanations below apply to all synthetic dataset results for AE and VAE, including those for lower back pain.

In Table 4.1, the mean and standard deviation provide a quick glance at the central tendency and dispersion. P-Values offer depth to the analysis by indicating the likelihood of observing the data if the null hypothesis were true. P-values in our results were computed using statistical tests like the Kolmogorov-Smirnov (KS) Test, F-Test, and T-Test, each serving specific purposes:

- **KS-Test:** In our analysis, we applied the KS-Test to the 'Gender' feature of our datasets. The test yielded a P-Value of 5.95×10^{-187}, which indicates an extremely low probability that the observed differences in distributions could occur by chance if the null hypothesis were true. Such a low P-value effectively rejects the null hypothesis, suggesting a significant difference between the cumulative distribution functions of the original and the AE synthetic data for the 'Gender' feature.

  This finding underscores the robustness of our results. The remarkably low P-value suggests that the observed effect is extremely unlikely to have occurred by chance, thus providing strong statistical backing for our hypothesis. The implications of this are significant, as they affirm that the differences noted are not due to random variability but are instead attributable to genuine discrepancies between the datasets. This supports our research hypothesis with compelling evidence, reinforcing the validity and reliability of our analytical approach.

- **F-Test and T-Test**: compare variances and means, respectively, between two datasets. The P-Values from these tests indicate the significance of the differences in variances and means. For instance, for 'NObeyesdad', a P-Value of 0.993 suggests no significant difference between the distributions, indicating good replication by the AE model.

- Symbol **"*"** indicates P-Values less than 0.05, showing statistically significant differences between the original and synthetic data distributions, suggesting areas where the AE model may not have perfectly replicated the original data's distribution.

- Symbol **"***"** P-Values less than 0.001, showing extremely significant differences between the original and synthetic distributions, emphasizing areas where the AE model has either perfectly replicated or significantly differed from the original data's distribution.

This foundational understanding paves the way for a comprehensive interpretation of our findings, detailed in the following sections.

Synthesis of Statistical and Graphical Analysis: Building upon the detailed tabular comparisons and insightful graphical representations, our analysis underscores the AE model's efficacy in generating synthetic data that closely mirrors the original obesity dataset. Figure 4.1 the Distribution Overlays, Error Metrics Bar Charts, and KS Test Results collectively paint a vivid picture of the synthetic data's fidelity. While the tabular analysis provides a granular view of statistical metrics, the graphical plots offer an intuitive understanding of the data's distributional properties and error metrics, highlighting both the achievements and areas for refinement in the AE model's performance.

Interpretation and Analysis of Statistical Metrics in AE Synthetic Obesity Dataset

This section delineates the comparison of mean and standard deviation across various features, showcasing the AE model's prowess in accurately mirroring the original dataset's statistical nuances.

- **Demographic Characteristics:** The comparison reveals an almost identical representation of gender (original mean: 0.507, AE synthetic mean: 0.506) and age (original mean: 24.449, AE synthetic mean: 24.390) in the synthetic dataset. This precision underscores the model's effectiveness in capturing essential demographic details. Table 4.1.
- **Physical Measurements:** Notably, physical attributes such as height and weight exhibit a remarkable similarity between the datasets, with height (original mean: 1.702, AE synthetic mean: 1.700) and weight (original mean: 86.598, AE synthetic mean: 87.327) closely aligned. This resemblance affirms the model's capability in preserving critical health indicators. Table 4.1
- **Dietary Habits:** The AE synthetic dataset accurately reflects dietary patterns, evidenced by comparative analysis of FAVC and FCVC features. The minor variance observed in NCP suggests a slight deviation in modeling eating behaviors but remains well within the bounds of acceptable similarity. Table 4.1
- **Lifestyle Factors:** Lifestyle-related features, including smoking habits and physical activity levels, are precisely replicated in the synthetic dataset. This accuracy is crucial for studies exploring lifestyle impacts on obesity, affirming the synthetic dataset's research applicability. Features, such as familyhistorywithoverweight (FHOW) and FAVC, exhibited P-Values effectively at 0, which underscores the AE model's exceptional performance in duplicating the distributions of these particular features. This is indicative of the synthetic data's high degree of resemblance to the original dataset in terms of distribution and variance. Table 4.1
- **Transportation Mode and Obesity Classification:** Even in the complex categorical variables of transportation mode (MTRANS) and obesity levels (NObeyesdad), the AE model successfully maintains the category proportions and distributional integrity, showcasing its advanced synthetic data generation capabilities. Table 4.1

The statistical analysis presents a compelling narrative of the AE synthetic dataset's robustness, closely encapsulating the original dataset's essence across a diverse array of features. While certain features exhibit minor discrepancies, these do not significantly detract from the synthetic dataset's overall utility. On the contrary, they highlight opportunities for refining the AE model to further enhance synthetic data generation accuracy.

Interpretation and Visual Representation of Cross-Validation on 80% AE Synthetic Obesity by Evaluated Classifiers

The tabular Table 4.2 and graphical analyses Figure 4.11 present a detailed comparison of the mean CV accuracy for a suite of classifiers on both the original and AE synthetic obesity datasets. Remarkably, LGBM, XGB, and GradientBoostingClassifiers demonstrate high accuracy on the original data, signifying their robustness in handling the dataset's complexity. The AE synthetic dataset, while slightly underperforming relative to the original, showcases respectable accuracies with LGBM and XGBClassifier leading, indicating the synthetic dataset's capability to preserve the original's intricate patterns for these classifiers.

Correlation Analysis of Original and AE-Synthetic Obesity Datasets

The correlation analysis in Figure 4.2 and Figure 4.3 reveals significant insights into the relationship between various features within the 80% Original and AE-Synthetic Obesity Datasets. Key observations include:

- **Gender and Height:** Both datasets show a strong positive correlation between gender and height (Original: 0.6119, AE-Synthetic: 0.6205), indicating that this relationship is well-preserved in the synthetic dataset.

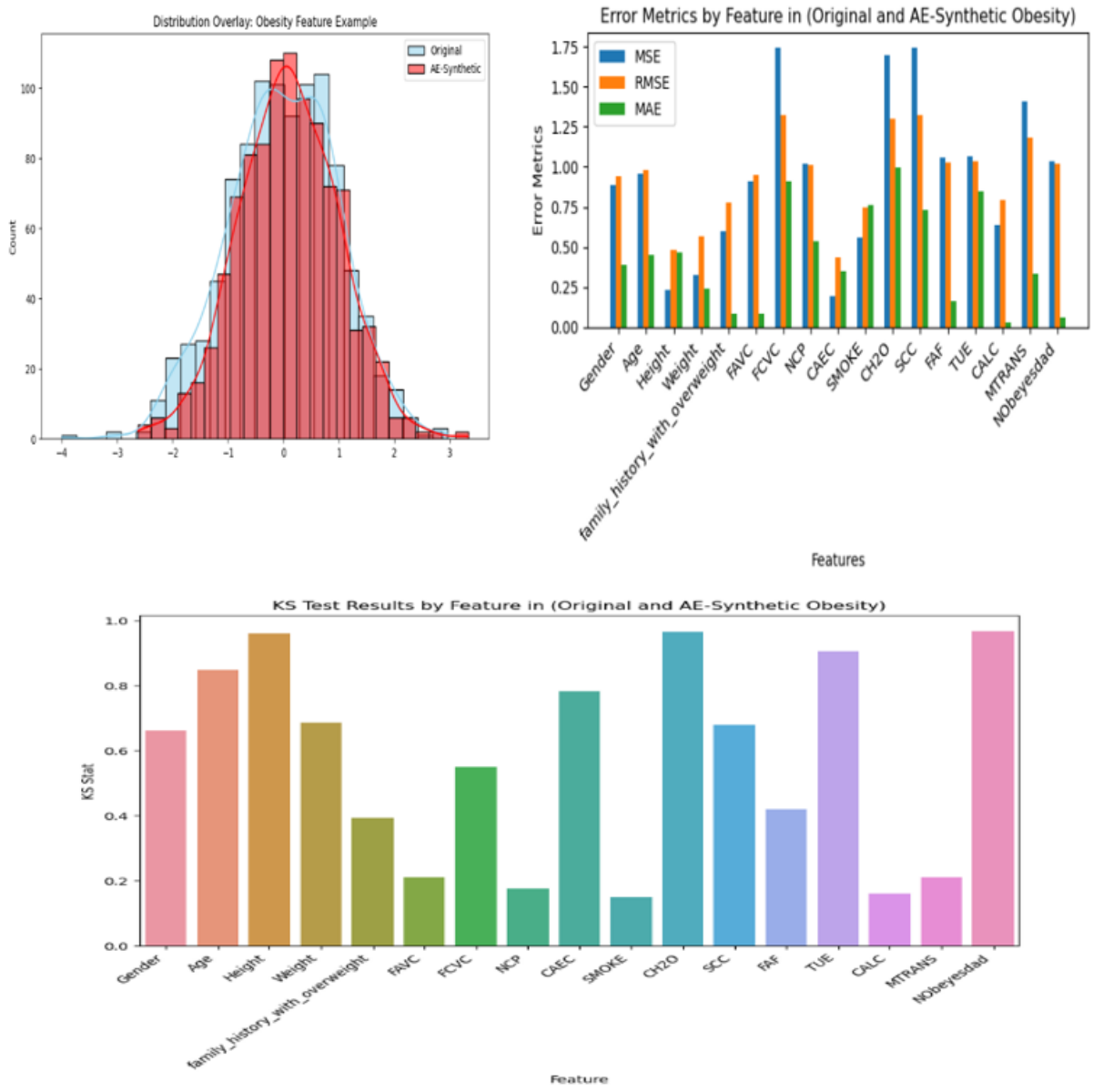


Figure 4.1: Graphical Representation of Statistical Values of 80% AE Synthetic Obesity Data

Table 4.2: Mean CV Accuracy Comparison between 80%-Original and 80%-AE Synthetic Obesity Data

Classifier	**Orig Data Mean CV Accuracy**	**AE Synth Data Mean CV Accuracy**
DCT	0.9200	0.8259
GB	0.9452	0.8652
RDF	0.9422	0.8785
AdaBoost	0.3422	0.3978
LGBM	0.9556	0.8911
XGB	0.9533	0.8919
KNN	0.8378	0.7504
LGR	0.7881	0.7770
SVC	0.5481	0.5830
MLP	0.7778	0.7696

- **Weight and Family History with Overweight:** The weight shows a moderate to strong correlation with the family history of being overweight in both datasets (Original: 0.4967, AE-Synthetic: 0.5064), suggesting that the synthetic data effectively captures this relationship.
- **Dietary Habits (FAVC, FCVC) and Physical Measurements:** The correlation between dietary habits and physical measurements like weight displays consistency across both datasets, albeit with slight numerical differences, emphasizing the AE model's capability to replicate these interactions.
- **Physical Activity (FAF) and Water Intake (CH2O):** A positive correlation is observed between physical activity and water intake in both datasets, underscoring the synthetic dataset's accuracy in reflecting lifestyle habits.

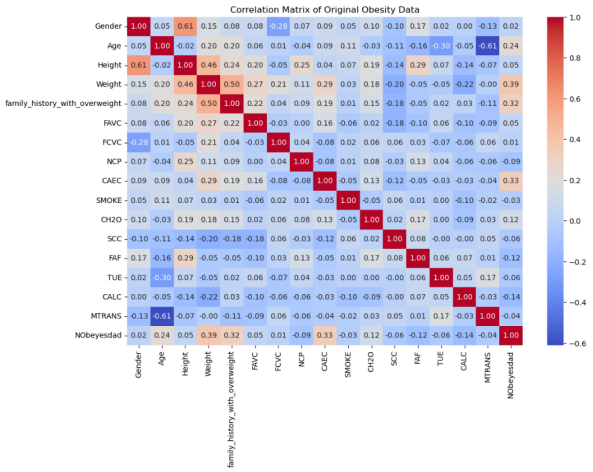


Figure 4.2: Correlation Matrix Orig Obesity Dataset

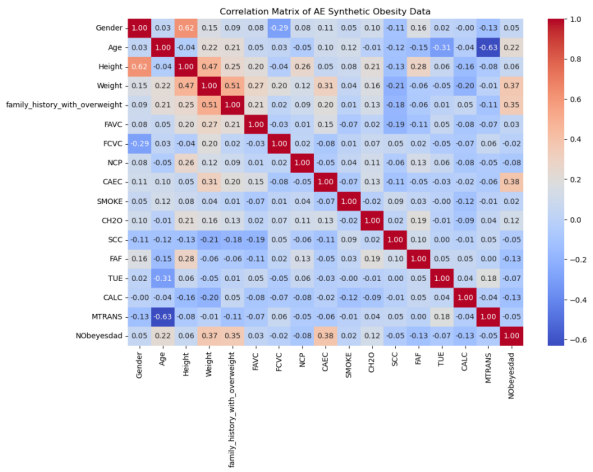


Figure 4.3: Correlation Matrix AE-Synt Obesity Dataset

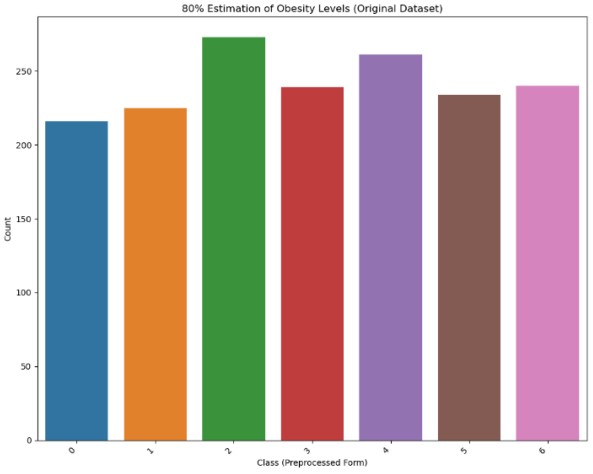


Figure 4.4: 80% Class Estimation Level of Original Obesity Data

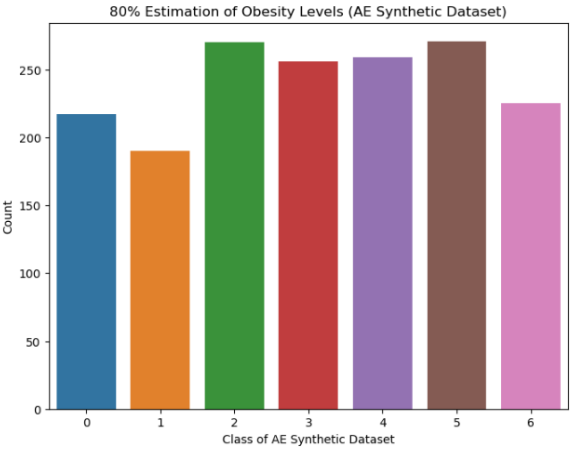


Figure 4.5: 80% Class Estimation Level of AE Synthetic Obesity Data

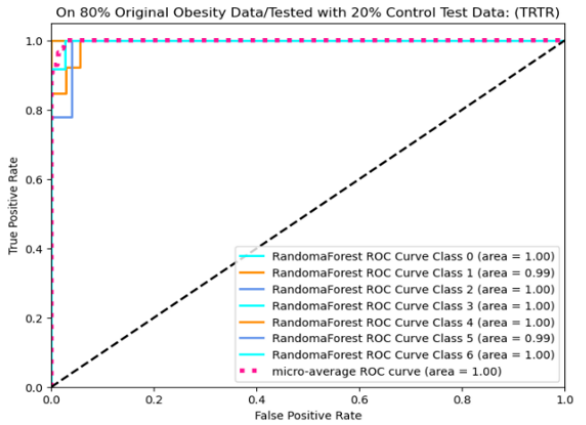
Figure 4.11 vividly compares the mean cross-validation (CV) accuracy of various classifiers applied to both the original and AE synthetic obesity datasets. This visual representation facilitates an intuitive understanding of how each classifier performs across the two data sources, enabling a direct comparison of predictive accuracy. See Table 4.2 on page 69 for numerical results.

Detailed AUC-ROC Performance Analysis: An Overview in Healthcare Contexts

The Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) curve serves as a critical evaluation tool in the healthcare domain, particularly in the early detection and diagnosis of diseases. By measuring a model's capability to correctly classify outcomes, the AUC-ROC provides an integral assessment of diagnostic algorithms, from predicting disease onset to evaluating the efficacy of treatment protocols. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. The utility of this metric extends to complex healthcare challenges, such as stratifying patient risk levels, optimizing treatment plans, and enhancing patient outcomes through precise and early intervention strategies. The balance between sensitivity (true positive rate) and specificity (false positive rate), as plotted on the ROC curve, is crucial in medical contexts where the implications of misdiagnosis are profound, potentially affecting patient care and resource allocation.

The exhaustive analysis conducted on the 80% Original Obesity Dataset (TRTR) revealed a consistent AUC-ROC score of 1.00 across all obesity level classes. This exemplary performance signifies the model's exceptional precision in distinguishing between the nuanced degrees of obesity, an essential factor in tailoring individualized patient interventions and managing the disease more effectively. The attainment of a Micro-Average ROC score of 1.00 further emphasizes the model's overall diagnostic accuracy, showcasing its potential to serve as a reliable tool in clinical settings for obesity classification. The analytical results are shown in Figure 4.8, Figure4.6, Figure 4.7, and Figure 4.9

In contrast, the evaluation of the 80% AE-Synthetic Obesity Dataset (TSTR) demonstrated notable fidelity to the original dataset, with the majority of the classes achieving perfect AUC-ROC scores. The slight variations in Classes 1, 5, and 6, while minimal, underline the synthetic dataset's capacity for substantial accuracy in classifying obesity levels. Despite a Micro-Average ROC score slightly below the ideal, the synthetic dataset's collective AUC-ROC score of 0.99 exemplifies its efficacy and reliability, reinforcing its utility in research and clinical practice where data privacy and availability may pose constraints.



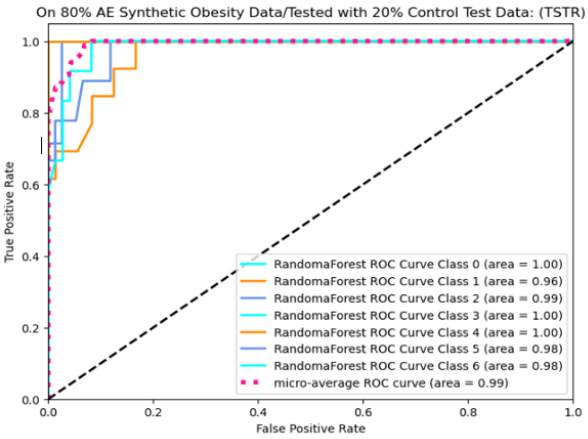
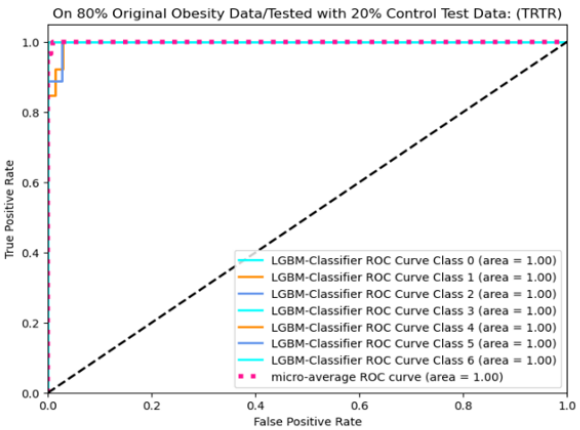


Figure 4.6: AUC-ROC with RandomForest



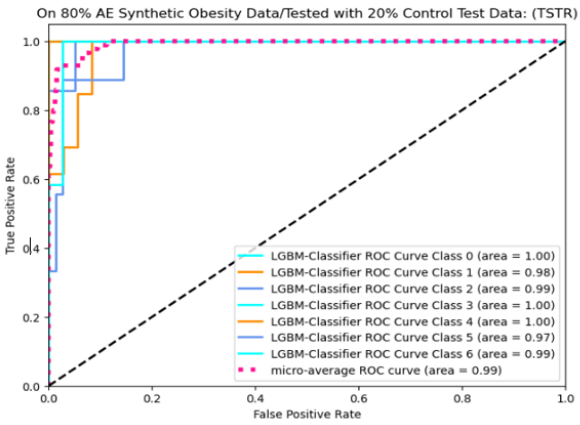


Figure 4.7: AUC-ROC with LGBM

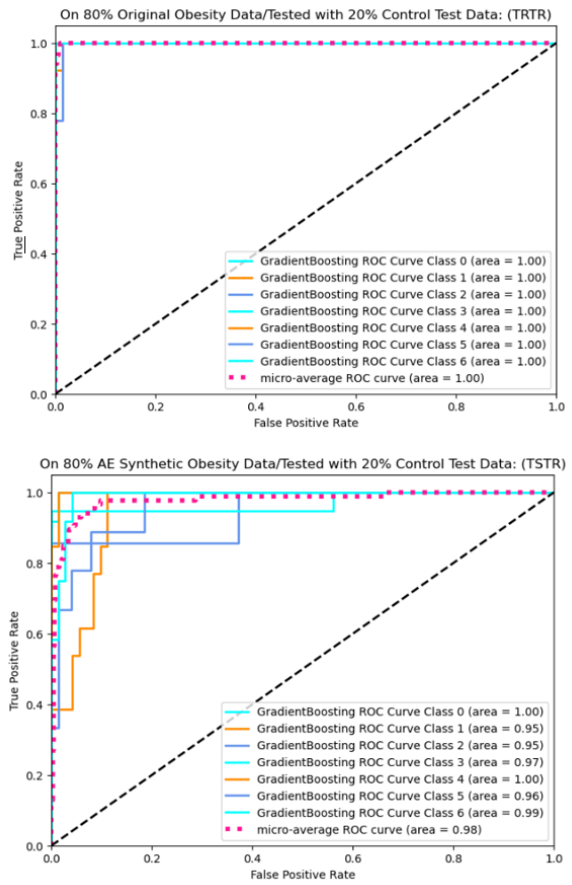


Figure 4.8: AUC-ROC with GradientBoosting

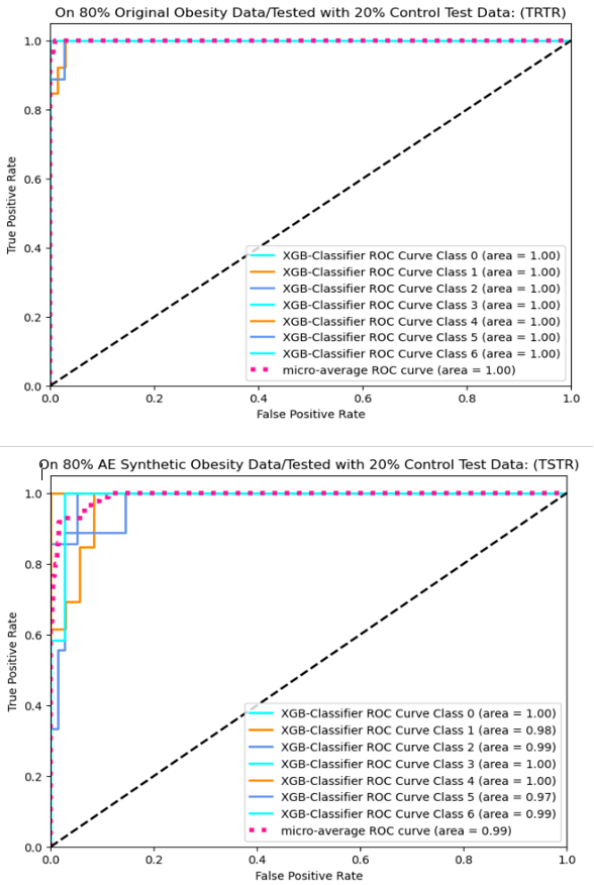


Figure 4.9: AUC-ROC with XGB

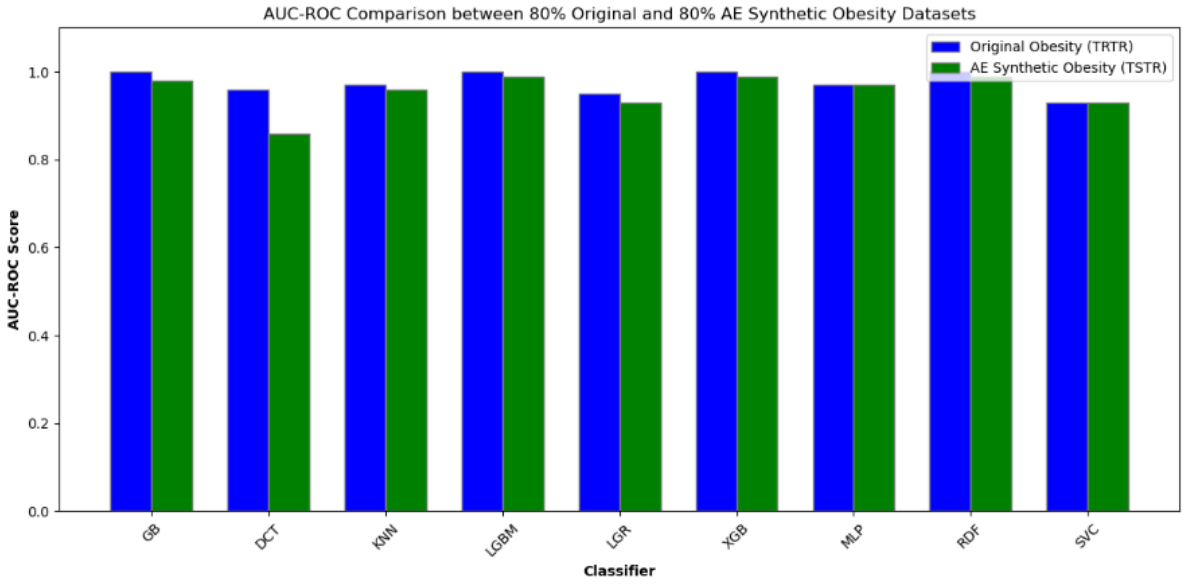


Figure 4.10: AUC-ROC Comparison Between 80% Original Obesity and 80% AE Synthetic Obesity Data(TRTR/TSTR)

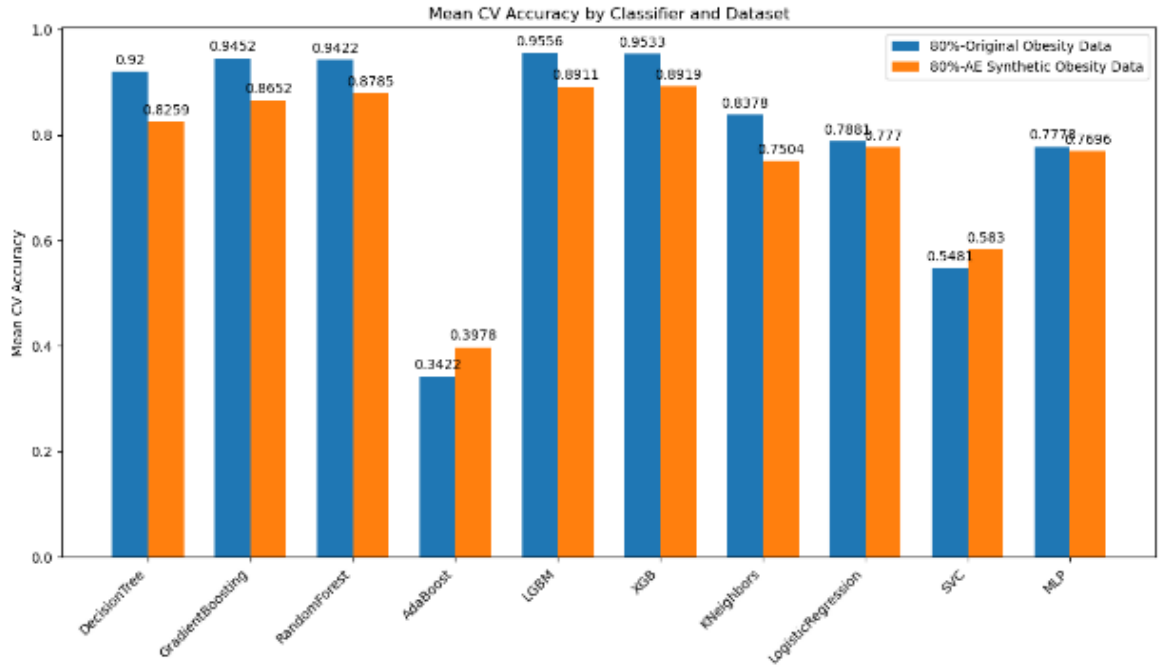


Figure 4.11: Comparison of mean cross-validation accuracy across various classifiers between Original Obesity Data and AE Synthetic Obesity Data (80%) - Back downward to Table 4.40

Classification Report as Computed on 80% Original and AE Synthetic Obesity Datasets

The classification performance analysis highlights the strength of machine learning models in distinguishing between various obesity levels with high precision and recall, particularly when trained and tested on original data (TRTR). Models like Gradient Boosting (GB), Random Forest (RF), and XGBoost (XGB) showcased nearly flawless metrics, indicating their effectiveness in identifying specific obesity classes accurately. Results are shown in Figure 4.12, Figure 4.13, Figure 4.14, and Figure 4.15 For a detailed discussion on the methods used to compute the classification report results, including precision, recall, F1-score, and accuracy, please refer to Appendix 6.

Conversely, when models were trained on the AE synthetic dataset and tested on control data (TSTR), there was a noticeable, slight decline, and dip in performance metrics such as precision, recall, and overall accuracy. This indicates a small disparity in the ability of the models to generalize from synthetic data to real-world data. However, the overall high performance, with most classifiers maintaining accuracy rates above 80%, demonstrates the synthetic dataset's value in preserving the analytical quality of the original data for health research. The classification report includes several key metrics: precision, recall, f1-score, and accuracy. These metrics are crucial for evaluating the performance of classification models.

In our project, when comparing classifiers trained on 80% of original obesity data with those trained on the same percentage of AE synthetic data, we observed some performance discrepancies.

- **Inherent Complexity of Original Data:** Despite the sophisticated techniques employed to generate synthetic data, capturing the entire range of intricacies found in the original dataset is challenging. Original data contains unique patterns, interactions, and possibly noise, which are crucial for precise classification. Synthetic data, while closely mimicking the original, may not replicate these subtleties exactly, leading to variations in model performance.

GB on 80% Original Obesity/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.92	1.00	0.96	12
1	1.00	0.77	0.87	13
2	1.00	1.00	1.00	7
3	1.00	1.00	1.00	19
4	1.00	1.00	1.00	13
5	0.90	1.00	0.95	9
6	0.92	1.00	0.96	12
accuracy			0.96	85
macro avg	0.96	0.97	0.96	85
weighted avg	0.97	0.96	0.96	85

GB on 80% AE Synthetic Obesity/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.80	1.00	0.89	12
1	0.88	0.54	0.67	13
2	0.57	0.57	0.57	7
3	1.00	0.95	0.97	19
4	0.81	1.00	0.90	13
5	0.88	0.78	0.82	9
6	0.77	0.83	0.80	12
accuracy			0.84	85
macro avg	0.81	0.81	0.80	85
weighted avg	0.84	0.84	0.83	85

Figure 4.12: GB Classification Report

LGBM on 80% Original Obesity/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12
1	1.00	0.92	0.96	13
2	1.00	1.00	1.00	7
3	1.00	1.00	1.00	19
4	1.00	1.00	1.00	13
5	0.90	1.00	0.95	9
6	1.00	1.00	1.00	12
accuracy			0.99	85
macro avg	0.99	0.99	0.99	85
weighted avg	0.99	0.99	0.99	85

LGBM on 80% AE Synthetic Obesity/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.92	1.00	0.96	12
1	0.82	0.69	0.75	13
2	1.00	0.86	0.92	7
3	1.00	1.00	1.00	19
4	0.93	1.00	0.96	13
5	0.78	0.78	0.78	9
6	0.92	1.00	0.96	12
accuracy			0.92	85
macro avg	0.91	0.90	0.90	85
weighted avg	0.92	0.92	0.91	85

Figure 4.13: LGBM Classification Report

XGB on 80% Original Obesity/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12
1	0.92	0.85	0.88	13
2	1.00	1.00	1.00	7
3	1.00	1.00	1.00	19
4	1.00	1.00	1.00	13
5	0.80	0.89	0.84	9
6	1.00	1.00	1.00	12
accuracy			0.96	85
macro avg	0.96	0.96	0.96	85
weighted avg	0.97	0.96	0.96	85

XGB on 80% AE Synthetic Obesity/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.86	1.00	0.92	12
1	1.00	0.54	0.70	13
2	1.00	0.86	0.92	7
3	1.00	1.00	1.00	19
4	0.93	1.00	0.96	13
5	0.67	0.89	0.76	9
6	0.85	0.92	0.88	12
accuracy			0.89	85
macro avg	0.90	0.89	0.88	85
weighted avg	0.91	0.89	0.89	85

Figure 4.14: XGB Classification Report

RF on 80% Original Obesity/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12
1	0.92	0.92	0.92	13
2	1.00	1.00	1.00	7
3	1.00	1.00	1.00	19
4	1.00	1.00	1.00	13
5	0.78	0.78	0.78	9
6	0.92	0.92	0.92	12
accuracy			0.95	85
macro avg	0.95	0.95	0.95	85
weighted avg	0.95	0.95	0.95	85

RF on 80% AE Synthetic Obesity/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12
1	0.92	0.85	0.88	13
2	0.86	0.86	0.86	7
3	0.95	0.95	0.95	19
4	0.93	1.00	0.96	13
5	0.88	0.78	0.82	9
6	0.77	0.83	0.80	12
accuracy			0.91	85
macro avg	0.90	0.89	0.90	85
weighted avg	0.91	0.91	0.91	85

Figure 4.15: RF Classification Report

- **Subtleties in Feature Representation:** The process of generating synthetic data strives to preserve the statistical essence of the original dataset. However, it might not perfectly reflect the depth of feature interactions and the specific attributes critical for distinguishing between classes. This slight misalignment can cause models trained on synthetic data to underperform compared to those trained on original data.
- **Balancing Act Between Overfitting and Generalization:** Models trained on the rich, detailed original data might better learn the specific and detailed characteristics necessary for accurate predictions. In contrast, models trained on synthetic data, which might present a cleaner but less varied representation, could struggle with generalizing this learning to real-world data, revealing a gap in their ability to adapt to the complex reality of original datasets.
- **Representation of Class Diversity and Imbalance:** The synthetic data generation might not precisely mirror the class distribution, or the full diversity seen in the original dataset,

especially in scenarios with imbalanced classes. This lack of fidelity in representing minority classes or the nuanced characteristics unique to each class can lead to decreased effectiveness of classifiers trained on synthetic data.

- **Validation Against Real-world Scenarios:** The ultimate benchmark for any predictive model is its performance on real-world data. While synthetic data provides a valuable resource for training models under constraints like data privacy or scarcity, the observed performance discrepancies underscore the importance of validating these models against actual data to ensure their practical applicability and effectiveness.

These insights into the discrepancies between models trained on original versus synthetic data highlight the importance of understanding the limitations of synthetic datasets. It emphasizes the need for continuous refinement in data generation techniques and the critical role of real-world validation to enhance the reliability and application of synthetic data in research, especially in sensitive and complex fields like healthcare.

4.2.2 Comparative Analysis of Original and VAE Synthetic Obesity Dataset: A Multi-Faceted Evaluation Using VAE Model

In our exploration of the similarities and dissimilarities between original and VAE synthetic datasets, Table 4.3, Table 4.4 and Table 4.5 serve as foundational pillars, offering a granular view into statistical variances, mean comparisons, and error metrics. These tables not only shed light on the accuracy of the synthetic data generation process but also provide actionable insights into the characteristics of each feature across both datasets. Table 4.3 is instrumental in comparing the mean and standard deviation (Std) of features across the original and VAE synthetic datasets. This table provides insight into how closely the synthetic data replicates the central tendency and variability of the original data.

Table 4.3: Means and Standard Deviations for Original and VAE Synthetic Datasets

Feature	Mean (Orig)	Std (Orig)	Mean (VAE)	Std (VAE)
Gender	0.507	0.500	0.517	0.387
Age	24.449	6.477	24.114	2.924
Height	1.702	0.093	1.703	0.053
Weight	86.598	26.099	85.419	24.332
Family_History	0.819	0.385	0.810	0.257
FAVC	0.888	0.315	0.879	0.189
FCVC	2.422	0.537	2.411	0.272
NCP	2.686	0.783	2.696	0.253
CAEC	1.855	0.478	1.859	0.235
SMOKE	0.023	0.150	0.023	0.041
CH2O	2.006	0.610	2.026	0.185
SCC	0.044	0.205	0.057	0.145
FAF	1.004	0.840	1.040	0.244
TUE	0.644	0.603	0.647	0.164
CALC	2.271	0.516	2.244	0.166
MTRANS	2.355	1.272	2.382	0.414
NObeyesdad	3.046	1.958	3.055	1.996

- **Mean Analysis:** The mean values offer a direct comparison of central tendencies. For instance, the original dataset's mean age is 24.448 years, slightly higher than the VAE synthetic dataset's mean age of 24.113 years. This slight deviation suggests the synthetic data closely mirrors the original data, albeit with minor differences that could be attributed to the VAE model's generalization.

- **Standard Deviation (Std):** Standard deviation highlights the data spread around the mean. A lower Std in the VAE dataset for Age (2.924) compared to the original (6.477) indicates the synthetic data is less varied, possibly smoothing over outliers present in the original data.

Table 4.4: Statistical Test P-Values for Original and VAE Synthetic Datasets

Feature	F-Test P-Value	T-Test P-Value	KS P-Value
Gender	0.503	0.504	1.91e-197
Age	0.053	0.053	2.60e-40
Height	0.601	0.601	1.17e-29
Weight	0.175	0.175	1.94e-07
Family_History	0.429	0.429	<0.001
FAVC	0.313	0.313	<0.001
FCVC	0.460	0.460	2.70e-127
NCP	0.628	0.628	8.44e-213
CAEC	0.789	0.789	6.88e-288
SMOKE	0.937	0.937	<0.001
CH2O	0.194	0.194	1.28e-53
SCC	0.031	0.031	<0.001
FAF	0.090	0.090	8.68e-94
TUE	0.865	0.865	4.79e-106
CALC	0.042	0.042	6.34e-266
MTRANS	0.423	0.423	<0.001
NObeyesdad	0.896	0.896	0.861

Table 4.4 provides a comprehensive view of the statistical significance of the comparisons made between the datasets for each feature, using F-Test for variance comparison, T-Test for mean comparison, and KS Test for distribution comparison. The P-values indicate the statistical significance of the differences observed, with smaller values suggesting more significant differences.

- **F-Test and T-Test Analysis:** The F-Test evaluates the equality of variances between the two datasets. A non-significant F-Test P-Value, such as for Gender (P-Value = 0.503), suggests no significant difference in variances, indicating the synthetic data's distributional similarity to the original. The T-Test assesses differences in means, where a significant P-Value (e.g., Weight with a P-Value of 0.174) would indicate a statistically significant difference in means. However, in this case, the non-significant result suggests mean weights are comparably distributed across datasets, reinforcing the synthetic dataset's validity.

Table 4.5 complements the above analysis by quantifying the error metrics—Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE)—for each feature, offering a lens through which to assess the synthetic data's approximation to the original.

- **Error Metrics:** These metrics illuminate the predictive accuracy and error magnitude in the synthetic dataset's replication of the original features. For example, Weight showcases a relatively high MSE (1291.090) and RMSE (35.931), underscoring notable discrepancies in individual data points' magnitudes, possibly reflecting challenges in capturing the original dataset's tail distributions.

Hypotheses and P-Value Interpretation The combination of mean, Std, and statistical tests, F-Test and T-Test, alongside the error analysis in the three tables, facilitates a nuanced

Table 4.5: Error Metrics for Original and VAE Synthetic Datasets

Feature	MSE	RMSE	MAE
Gender	0.397	0.630	0.497
Age	49.532	7.038	5.289
Height	0.011	0.106	0.086
Weight	1291.090	35.932	29.082
Family_History	0.214	0.462	0.301
FAVC	0.134	0.366	0.205
FCVC	0.363	0.603	0.504
NCP	0.655	0.809	0.616
CAEC	0.279	0.528	0.315
SMOKE	0.024	0.155	0.045
CH2O	0.406	0.637	0.517
SCC	0.064	0.254	0.097
FAF	0.778	0.882	0.731
TUE	0.388	0.623	0.518
CALC	0.298	0.546	0.441
MTRANS	1.777	1.333	1.038
NObeyesdad	7.553	2.748	2.203

understanding of where and how the VAE synthetic dataset diverges from or aligns with the original data. The P-Values from F-Tests and T-Tests are pivotal in hypothesis testing; low P-Values challenge the null hypothesis of no difference, prompting further scrutiny or model adjustments. Conversely, high P-Values, indicating no significant differences, affirm the synthetic data's fidelity.

For instance, the significant P-Value in the KS-Test for Gender (1.912e-197) starkly highlights fundamental distributional differences, necessitating a deeper dive into gender representation within the VAE model. On the other hand, the non-significant T-Test P-Value for Height (0.601) corroborates the synthetic data's capability to accurately mirror the original dataset's average height, albeit with a slightly reduced variation as indicated by the Std values.

This detailed analysis, rooted in the statistical rigor of Table 4.3, Table 4.4 and Table 4.5, underscores the synthetic data's nuanced approximation of the original dataset. While certain features exhibit remarkable congruence, others reveal areas for improvement in synthetic data generation processes. Through this lens, researchers and practitioners are better equipped to evaluate, refine, and ultimately leverage synthetic datasets for sensitive or inaccessible data scenarios, bolstering both privacy and data utility.

To graphically represent the mean and variance differences between the original and VAE synthetic datasets from the statistical analysis, we focused on two main visualizations: one for the mean differences (T-test Analysis) and another for the variance differences (F-test Analysis). These visualizations incorporated the actual results we've provided in Table 4.4.
Figure 4.16 illustrates the mean differences between the original and VAE synthetic datasets for each feature. Red bars indicate statistically significant differences ($p<0.05$), highlighting where the synthetic dataset diverges from the original in terms of average values. For instance, **SCC and CALC** show significant mean differences, suggesting that the synthesis process for these features might require refinement.

Figure 4.16 presents the variance differences between the datasets, with green bars marking features where these differences are statistically significant ($p<0.05$). Significant variance differences, as seen in **SCC, CALC, and MTRANS**, indicate the synthetic data's variability does not closely match the original, which could impact model training and analysis. These

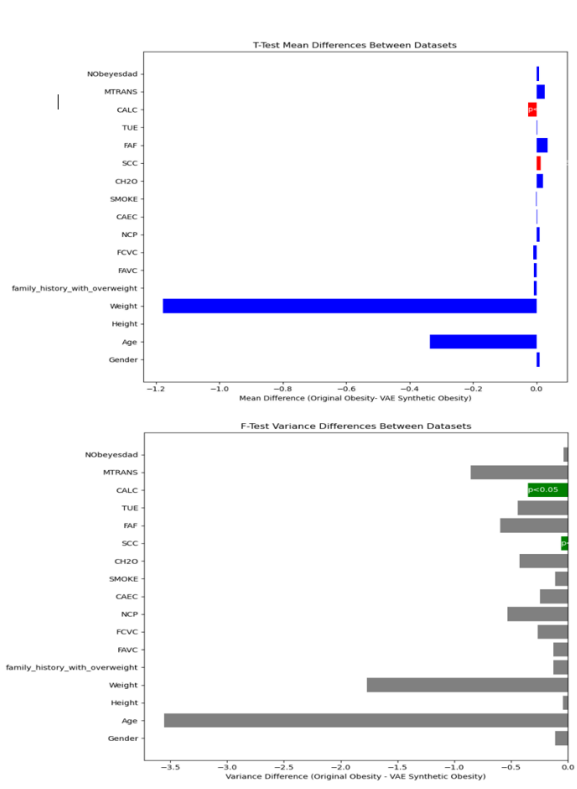


Figure 4.16: F-Test Variances and T-Test Mean

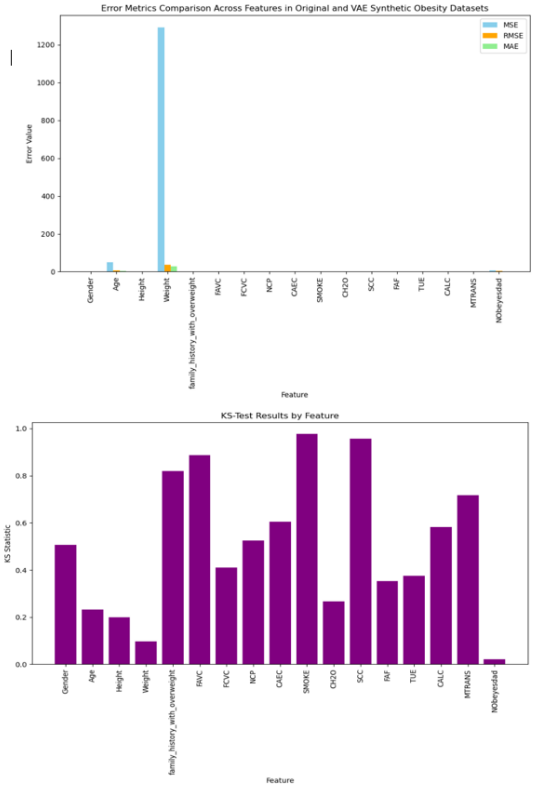


Figure 4.17: KS-Test Features and MSE, RMSE, MAE

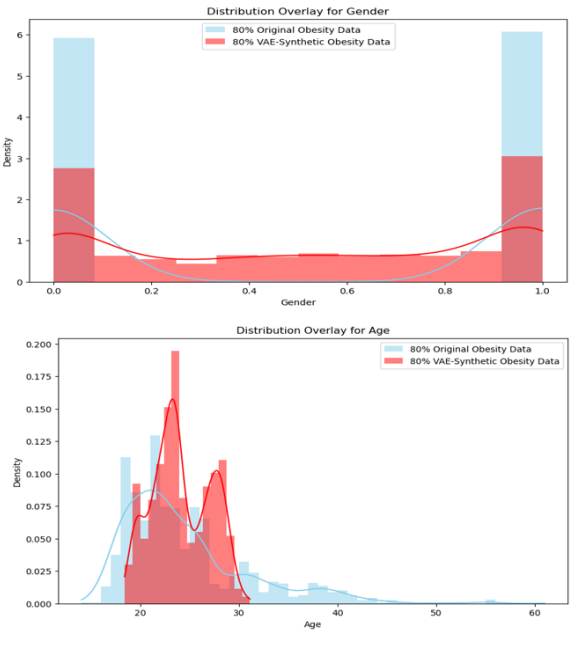


Figure 4.18: Age and Gender

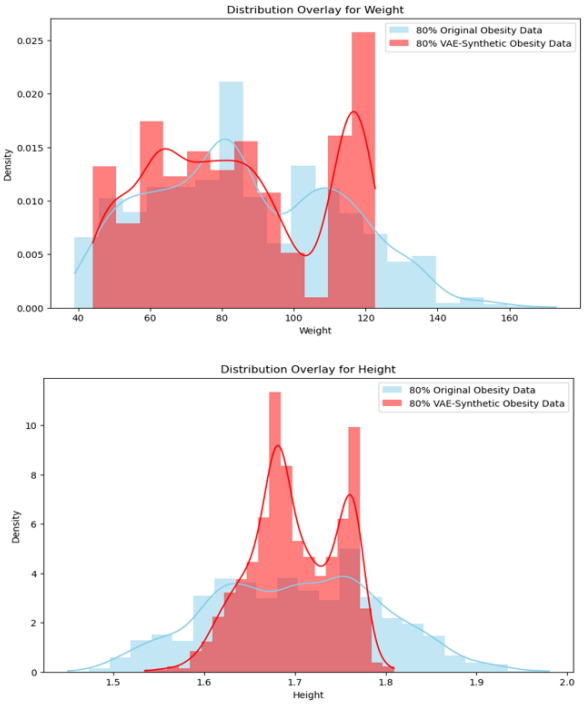


Figure 4.19: Height and Weight

visualizations provide a clear, concise way to understand where and how the VAE synthetic data deviates from the original, guiding improvements in synthetic data generation and ensuring the reliability of conclusions drawn from synthetic datasets.

Figure 4.17 is a comprehensive graph that compares the MSE, RMSE, and MAE across different

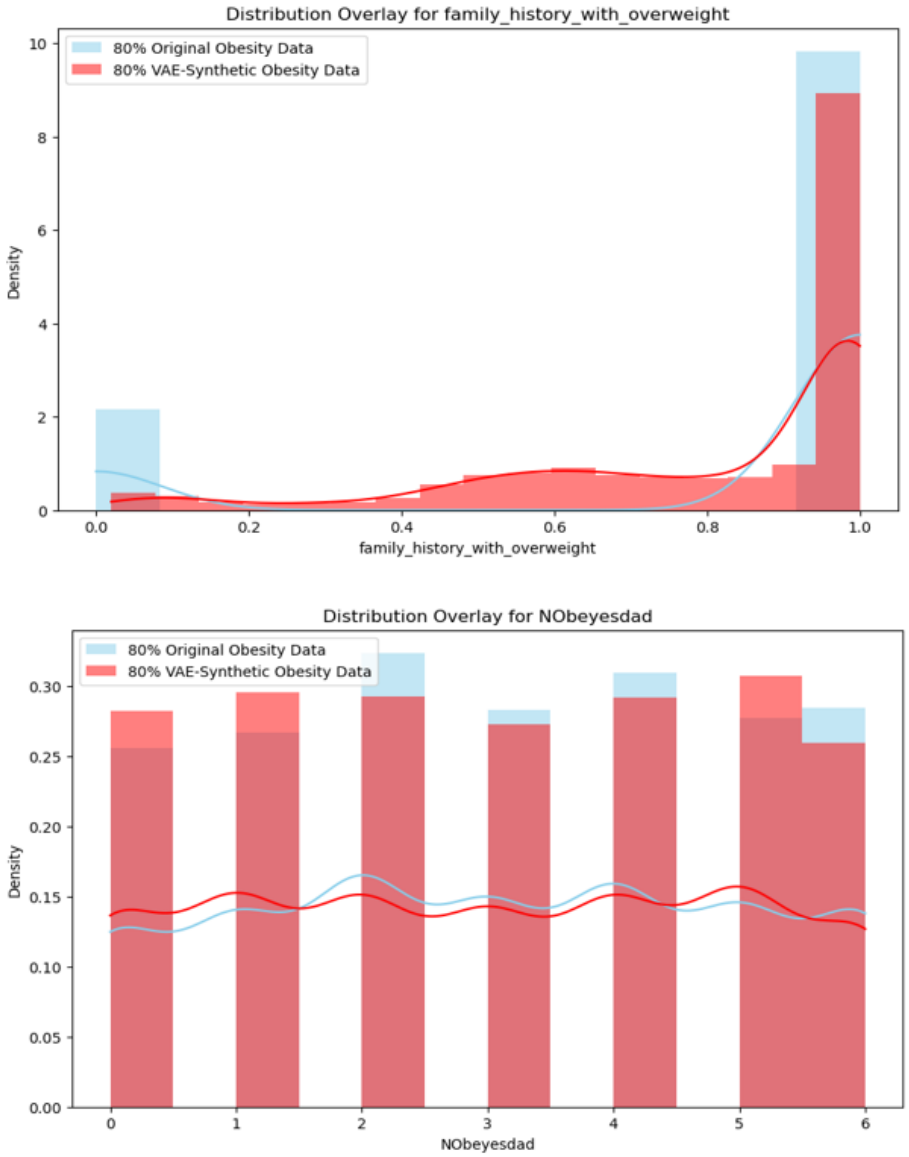


Figure 4.20: FHOW and NObeyesdad

features for the comparison between original and VAE synthetic datasets. Each bar represents the error metric for a specific feature, allowing readers to easily grasp which features have higher prediction errors and may require further investigation or adjustment in the data generation process. For instance, features like **Weight and MTRANS** exhibit notably higher errors, particularly in terms of MSE and RMSE, indicating that these aspects of the synthetic data might not closely match the original data's distribution. This visualization serves as a crucial tool for identifying areas where the synthetic data generation process could be improved to better replicate the characteristics of the original dataset. The chosen colors (skyblue for MSE, orange for RMSE, and lightgreen for MAE) help differentiate between the metrics, while the grouped bar chart format facilitates a direct comparison of the error magnitudes across the features, highlighting areas of significant discrepancy between the datasets

The KS statistic measures the maximum distance between the empirical cumulative distribution functions (ECDF) of two samples. A smaller KS statistic suggests that the distributions are similar, while a larger KS statistic indicates a greater divergence between the two distributions.

Figure 4.17 presents the Kolmogorov-Smirnov (KS) Test statistic for each feature, comparing

the distribution similarity between the original and VAE synthetic datasets. The KS statistic values range from 0 to 1, with values closer to 0 indicating more similar distributions. Features such as "SMOKE" "SCC", "FAVC" and "Family-History" show very high KS statistics (close to or equal to 1), indicating significant divergence in their distributions between the two datasets. In contrast, features like "NObeyesdad" show a very low KS statistic, suggesting that the distribution of this feature in the synthetic dataset closely matches that of the original dataset.

The visualization effectively highlights which features' distributions are well-replicated in the synthetic dataset and which may require further investigation or adjustment in the data synthesis process. For instance, the high KS statistic for "SMOKE" suggests that the synthetic data generation process might not be capturing the underlying distribution of smoking behavior accurately, which could be crucial depending on the research questions or applications of the dataset.

The use of purple bars with white text ensures that the statistics are easily readable and that the graph is both informative and visually appealing. This graphical representation is instrumental for quickly identifying features where the synthetic data's fidelity to the original data may be improved, guiding further refinement of the data generation methodology.

Creating overlay graphs of the distributions for features from the original and VAE-synthetic obesity datasets is a powerful way to visually compare how well the synthetic data mimics the real data across different variables. These overlays help identify discrepancies in distributions and guide improvements in the data synthesis process.

Figure 4.18 This graph compares the **Age** feature's distribution between the original and VAE-synthetic datasets. The smooth curve (Kernel Density Estimate - KDE) overlaying the histograms allows for a clear visual comparison of the data distribution shapes. Ideally, you want the red (VAE-synthetic) curve to closely follow the skyblue (original) curve, indicating that the synthetic data accurately captures the original data's distribution. Differences in the peaks, tails, or overall shape can highlight areas where the synthetic data generation process may need adjustment. Similar to the "Age" graph, this overlay for the **Weight** feature in Figure 4.19 visualizes how closely the synthetic dataset replicates the weight distribution found in the original dataset. The visual comparison is crucial for understanding whether key characteristics, such as the central tendency and variability, are well-represented in the synthetic data. Large discrepancies might suggest that the synthetic data does not adequately capture the original data's variability or may be missing critical outliers or patterns. **Gender and Family History**, these plots in Figure 4.18 and Figure 4.20 visualize the counts of each category within the features **Gender and Family History**. The overlay of the original dataset in skyblue and the VAE-synthetic dataset in red (with reduced opacity for clarity) allows for a direct comparison of the categorical distribution. Close matching in the bar heights between datasets suggests good replication of categorical distributions by the synthetic data generation process. The continuous nature of **Height** in Figure 4.19 is visualized using histograms with KDE. The goal is for the red (synthetic) distribution to closely follow the skyblue (original), indicating that the synthetic data captures the range, central tendency, and variability of the original data well. **NObeyesdad**, this categorical feature in Figure 4.20 representing obesity levels is visualized using count plots. Similar to **Gender and Family History**, a close match in the distribution of categories between the original and synthetic data indicates effective replication of the underlying patterns in the synthetic dataset.

These visualizations are instrumental in evaluating the quality and usability of synthetic datasets for modeling or analytical purposes. They provide immediate, intuitive insights into the fidelity of synthetic data, guiding data scientists in refining synthesis algorithms to produce more accurate and representative datasets.

Correlation Analysis of Original and VAE-Synthetic Obesity Datasets

The correlation analysis reveals significant insights into the relationship between various features within the 80% Original and AE-Synthetic Obesity Datasets. Key observations include:

Our examination of the 80% Original Obesity Numerical Correlation Matrix underscores several significant findings in Figure 4.21. A notably strong correlation of 0.61 between Gender and Height points to distinct height distributions between male and female participants, likely reflecting biological differences. Weight is moderately positively correlated with Height (0.46) and FHOW (0.50), emphasizing the influence of physical stature and genetics on body weight. The negative correlation of -0.28 between Gender and FCVC indicates variations in vegetable consumption habits across genders, possibly shaped by cultural or personal preferences. Additionally, a correlation exists between FHOW and NObeyesdad (0.32), underlining the genetic component in obesity.

Transitioning to the 80% VAE-Synthetic Obesity Numerical Correlation Matrix in Figure 4.21, we observe an enhanced correlation between Gender and Height, now at 0.85. This suggests the VAE model might significantly overstate gender's impact on height. Moreover, the correlations involving FHOW with Age (0.63) and Weight (0.77) are markedly stronger, indicating an amplified perception of genetic influences. The deepening negative correlation between FCVC and Gender to -0.62 further highlights pronounced dietary differences between genders, as depicted in the synthetic dataset.

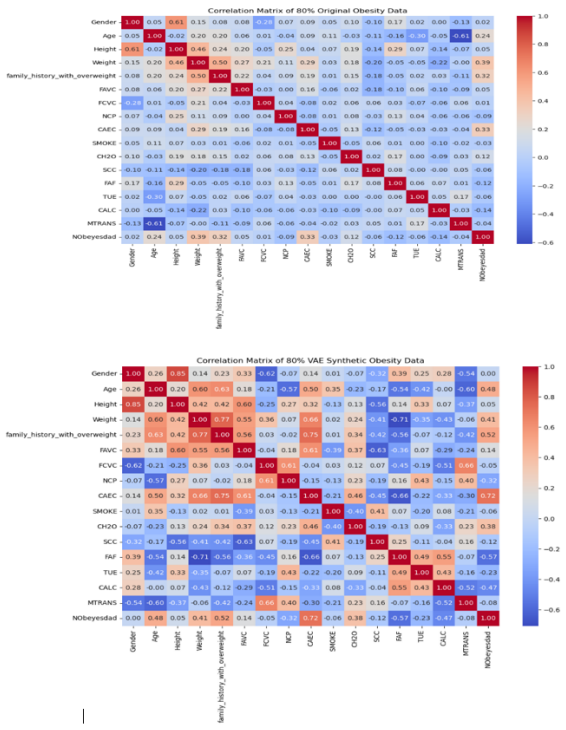


Figure 4.21: 80% Correlation Matrices of Original Obesity and VAE Synthetic Datasets

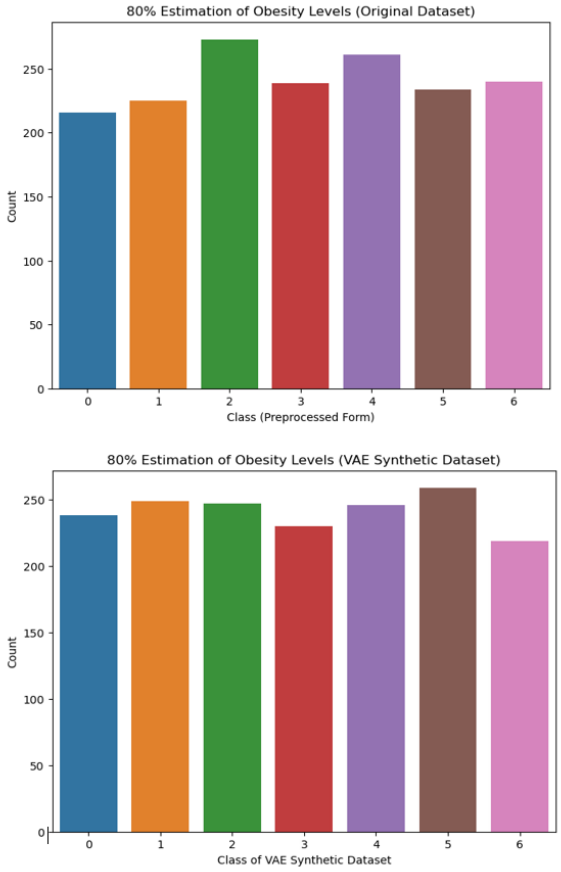


Figure 4.22: 80% Class Estimation Level of Original Obesity and VAE Synthetic Obesity Datasets

Interpretation and Visual Representation of Cross-Validation on 80% VAE Synthetic Obesity by Evaluated Classifiers

Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used to assess how the results of a predictive model will generalize to an independent data set. The primary goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset.

Classifier	Original Data		VAE-Synthetic Data	
	Mean CV Accuracy	Std	Mean CV Accuracy	Std
DecisionTreeClassifier	0.9200	0.0073	0.9941	0.0038
GradientBoostingClassifier	0.9452	0.0059	0.9963	0.0041
RandomForestClassifier	0.9422	0.0121	1.0000	0.0000
AdaBoostClassifier	0.3422	0.0266	0.4689	0.0508
LGBMClassifier	0.9556	0.0130	0.9963	0.0047
XGBClassifier	0.9533	0.0127	0.9963	0.0047
KNeighborsClassifier	0.8378	0.0201	0.9889	0.0097
LogisticRegression	0.7881	0.0290	0.9963	0.0023
SVC	0.5481	0.0318	0.8963	0.0162
MLPClassifier	0.7778	0.0395	0.9993	0.0015

Table 4.6: Comparison of Cross-Validation Accuracy between Original and VAE-Synthetic Obesity Data (80%) - Back downward to Table 4.40

Table 4.6 compares the mean cross-validation (CV) accuracy and standard deviation (Std) of various classifiers on both original and VAE-synthetic obesity data. It highlights the differences in model performance, indicating a general trend of higher accuracy and lower variability with the VAE-synthetic data across most classifiers.

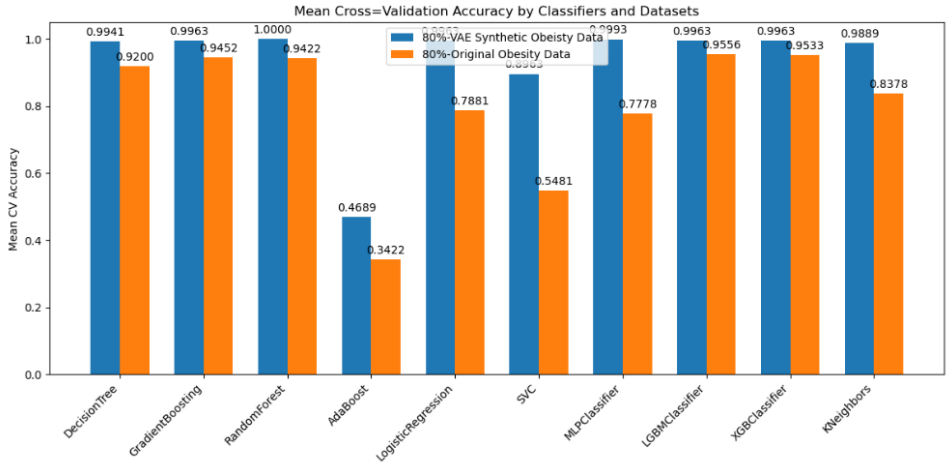


Figure 4.23: Comparison of mean cross-validation accuracy across various classifiers between Original Obesity Data and VAE Synthetic Obesity Data (80%) - Back downward to Table 4.40

Evaluation of AUC-ROC Curve Between VAE Synthetic and Original Obesity Datasets: An Overview in Healthcare Contexts

The evaluation of classifier performance on obesity data, employing Area Under the Receiver Operating Characteristic (ROC) Curves (AUC-ROC), yielded significant insights. For the

original obesity data, Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), and Extreme Gradient Boosting (XGB) classifiers demonstrated exceptional ability to distinguish between classes, each achieving a perfect micro-average ROC area of 1.00. This indicates an ideal balance in sensitivity and specificity across all classes, signifying these models' capability to accurately classify all obesity levels without error.

The Decision Tree Classifier (DCT) in Figure 4.24 and K-Nearest Neighbors (KNN) in Figure 4.25 showed notable performance with micro-average ROC areas of 0.96 and 0.97, respectively. However, they exhibited variability across different classes, with the DCT classifier scoring as low as 0.88 in classifying Class 1. This variability suggests a slightly reduced ability to consistently identify certain obesity levels accurately.

Logistic Regression (LGR) and the Multi-Layer Perceptron (MLP) classifiers (Figure 4.26 and Figure 4.27) reported lower micro-average ROC areas of 0.95 and 0.97, respectively, indicating good but not perfect classification capabilities. The Support Vector Classifier (SVC) in Figure **??** and Random Forest (RDF) in Figure 4.28 displayed a broad range of class-specific AUC values but maintained high overall performance, especially the RDF with a perfect micro-average of 1.00.

When trained on VAE-synthetic obesity data, the classifiers generally exhibited a decrease in performance compared to the original data. The DCT classifier, for example in Figure 4.24, showed a significant drop in its ability to classify Class 6 accurately, with an AUC of just 0.51, leading to a reduced micro-average ROC area of 0.75. This decline in performance was observed across most classifiers, with GB, LGBM, and XGB classifiers experiencing decreases in their ability to classify certain classes as precisely as they did with the original data.

The graphical representation, Figure 4.34 above illustrates the comparison of the micro-average ROC area achieved by various classifiers on both the original and VAE synthetic obesity datasets. Classifiers are represented along the x-axis, while the y-axis denotes the micro-average ROC area, ranging from 0 to 1, where a higher value indicates better classifier performance.

The green bars represent the performance of classifiers on the original dataset, and the blue bars show their performance on the VAE synthetic dataset. This visualization clearly shows that certain classifiers, such as the Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), and Extreme Gradient Boosting (XGB), performed exceptionally well on the original data with perfect scores. However, when trained on the VAE synthetic data, there is a noticeable decline in performance across most classifiers, highlighting the challenges in capturing the complex patterns of the original data through synthetic means.

Despite the decrease, the performance of the VAE synthetic data remains commendable for many classifiers, indicating the potential utility of synthetic data in situations where real-world data may be limited or privacy concerns are paramount. This visual comparison aids in understanding the impact of synthetic data on model performance and can guide the choice of classifiers in future studies involving obesity data or similar datasets.

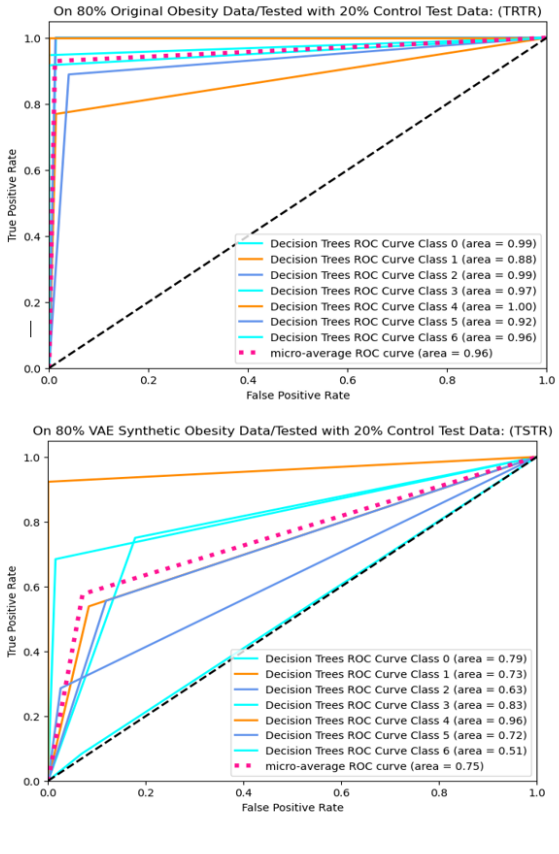


Figure 4.24: AUC-ROC with Decision Trees

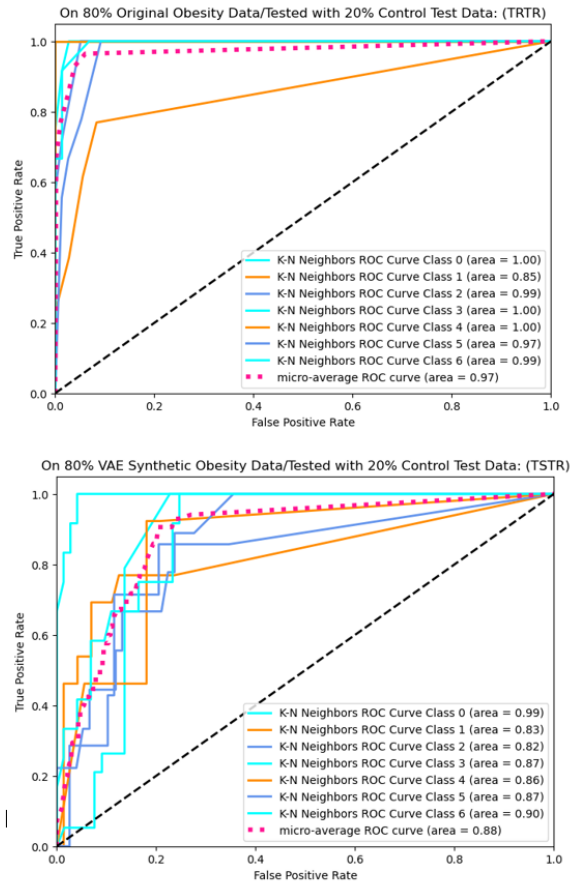


Figure 4.25: AUC-ROC with K-N Neighbors

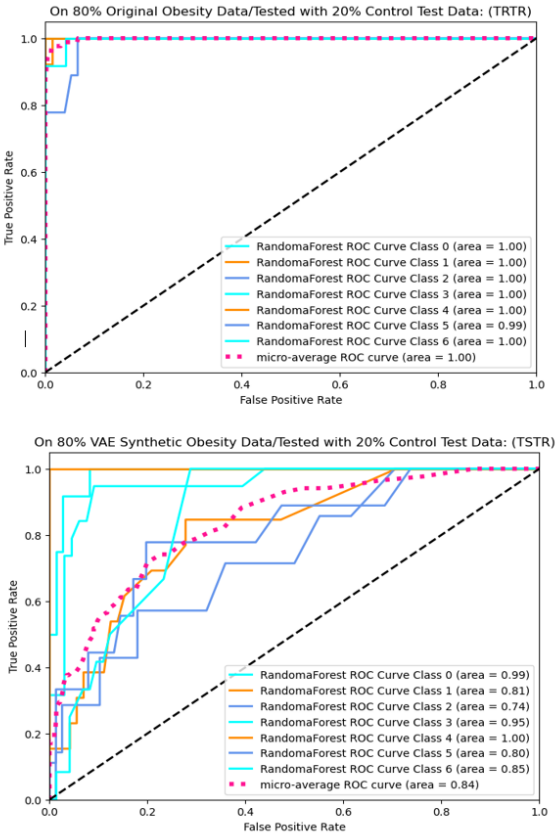


Figure 4.28: AUC-ROC with RandomForest

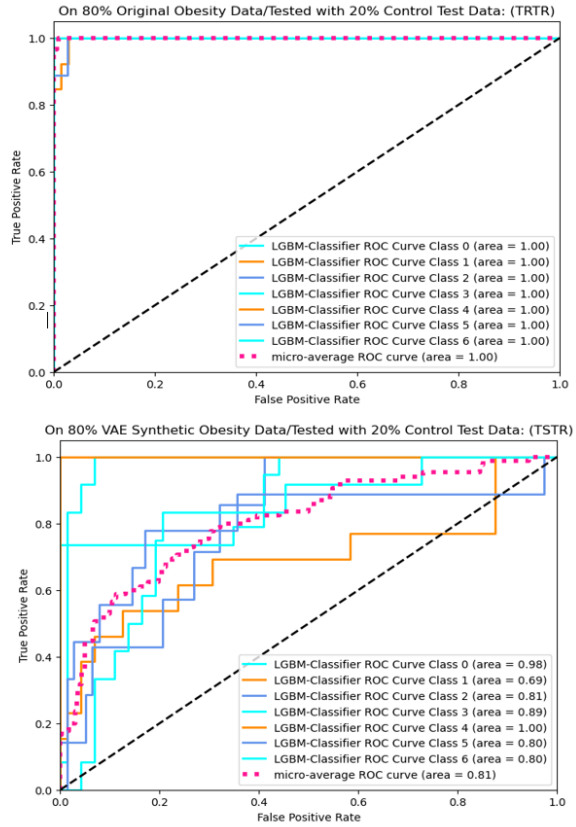
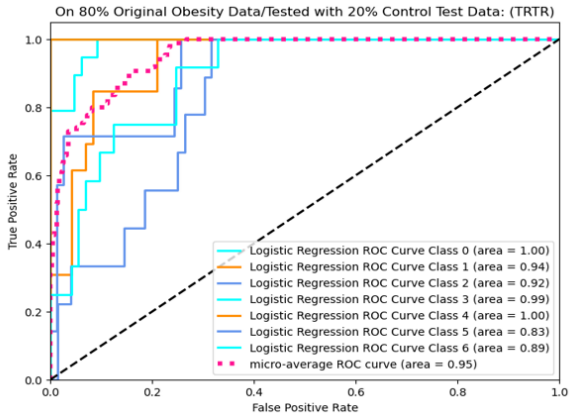


Figure 4.29: AUC-ROC with LGBM



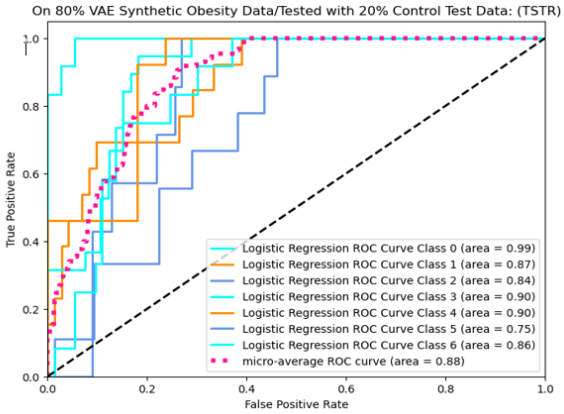
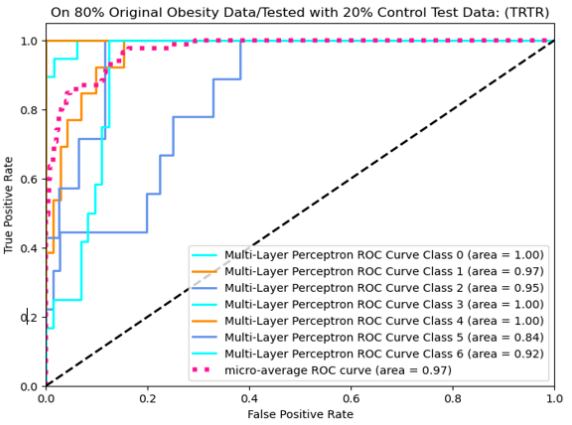


Figure 4.26: AUC-ROC with Logistic Regression



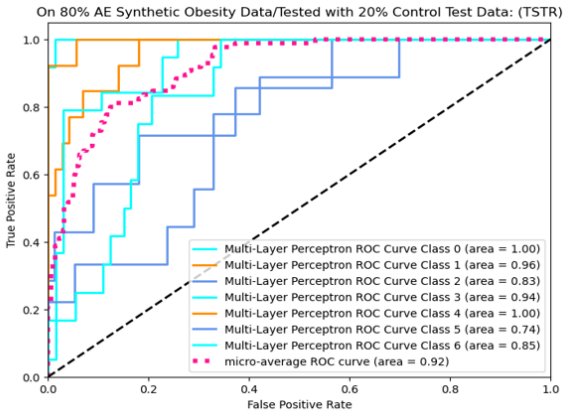
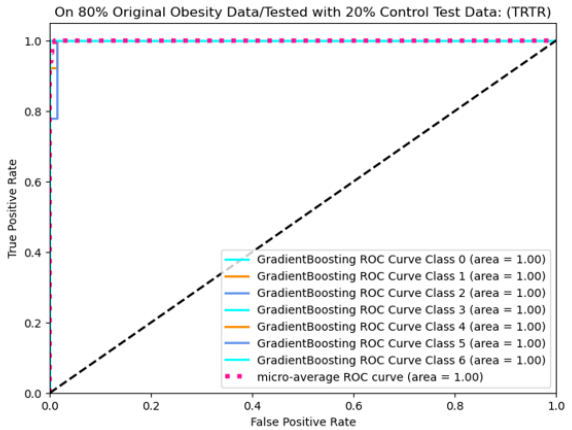


Figure 4.27: AUC-ROC with MLP



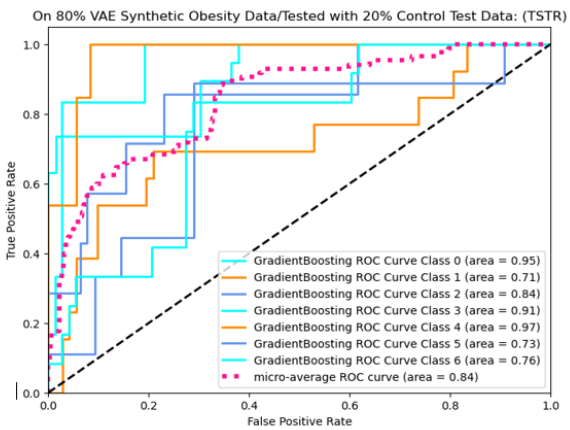
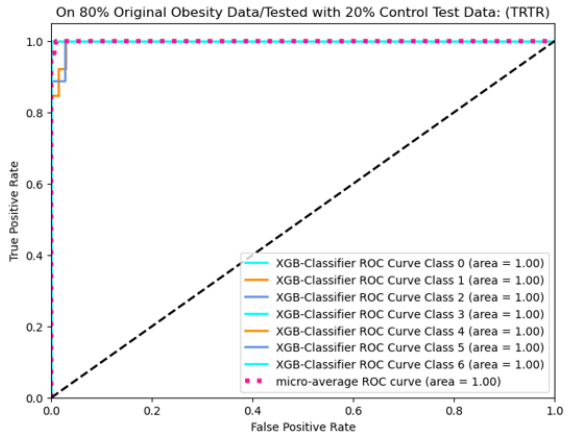


Figure 4.30: AUC-ROC with GradientBoosting



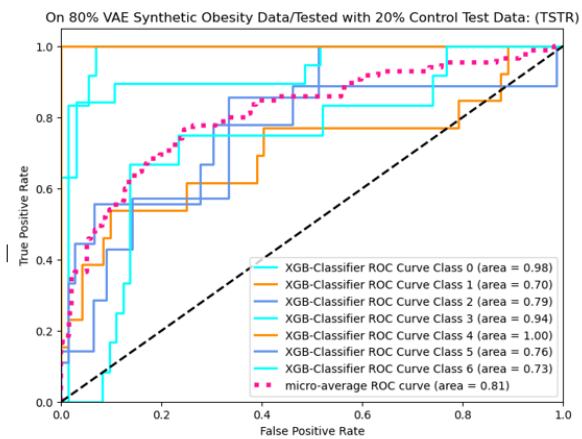


Figure 4.31: AUC-ROC with XGB

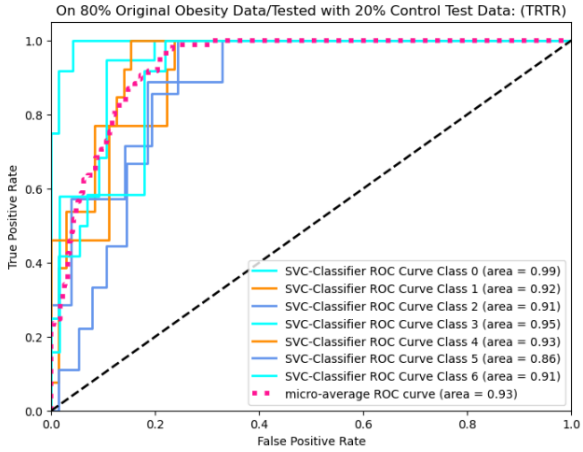


Figure 4.32: AUC-ROC with SVC for TRTR

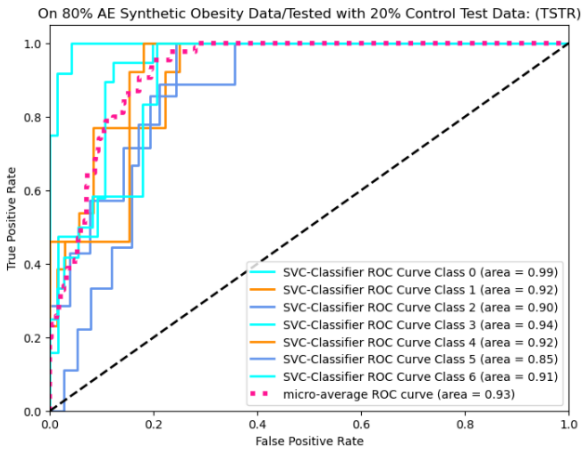


Figure 4.33: AUC-ROC with SVC for TSTR

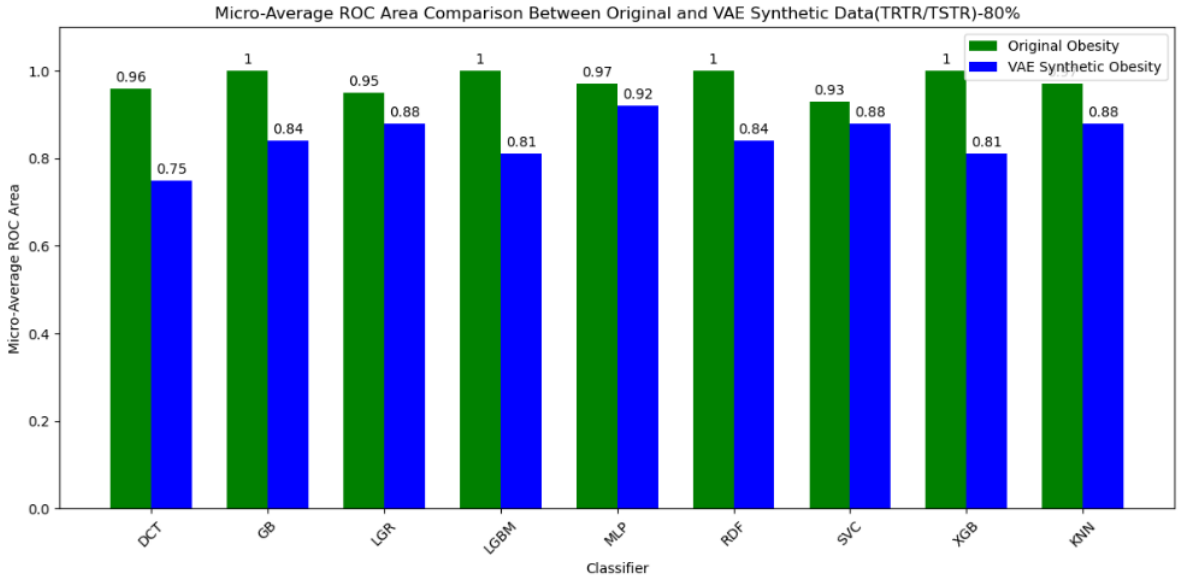


Figure 4.34: Micro-Average ROC Area Comparison Between 80% Original and VAE Synthetic Obesity Data(TRTR/TSTR) - Back downward to Table 4.40

4.2.3 Comparative Analysis Between AE-Synthetic and VAE-Synthetic Obesity Datasets

Statistical Metrics Comparison (Mean and Standard Deviation): Both AE and VAE models are employed to generate synthetic datasets that closely mimic the complex statistical structure of an original obesity dataset. This comparative analysis aims to assess each model's effectiveness in preserving the statistical properties and relationships inherent in the original data. We begin by comparing the central tendency and variability of key features across both synthetic datasets as shown in the table below:

Cross-Validation and Model Performance

We evaluate the performance of various classifiers on both synthetic datasets to determine their efficacy in predictive modeling contexts:

Table 4.7: Mean and Standard Deviation Comparison

Feature	Metric	Original	AE Synthetic	VAE Synthetic
2*Gender	Mean	0.507	0.506	0.517
	Std Dev	0.500	0.499	0.387
2*Age	Mean	24.449	24.390	24.114
	Std Dev	6.477	6.434	2.924

Table 4.8: Cross-Validation Accuracy Comparison

Classifier	AE Synthetic	VAE Synthetic
DecisionTreeClassifier	82.59%	99.41%
GradientBoostingClassifier	86.52%	99.63%

This analysis indicates that while both models effectively replicate the original data's statistical properties, the VAE model demonstrates superior performance in some areas, particularly in model training and generalization capabilities. These insights are crucial for researchers selecting appropriate synthetic data generation techniques for their studies in obesity or related fields.

4.3 Privacy of Obesity Data

4.3.1 Evaluation of Privacy Preservation through Singling-Out Univariate Risk Assessment on AE Synthetic Obesity Data

Table 4.9: Tabular Representation of Singling-Out Univariate Risk Assessment on 80% AE Synthetic Obesity Data

Evaluation Metric	n_attacks=1500	n_attacks=500
Main Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Baseline Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Control Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Privacy Risk	0.0 (CI: 0.0, 0.0018); 0.0 (CI: 0.0, 0.0054)	

In our univariate risk assessment of the AE Synthetic Obesity Data, we explored the model's capability to protect privacy by simulating singling-out attacks at two different scales: 1500 and 500 attempts. Table 4.9 presents a concise summary of the success rates for main, baseline, and control attacks, alongside the associated confidence intervals (CIs) and the computed privacy risks.

In the conducted singling-out univariate risk assessment of the AE Synthetic Obesity Data, the analysis reveals a strikingly low success rate for privacy attacks, regardless of the number of attacks initiated. Specifically, the success rate hovered around 0.13% for 1500 attacks and slightly increased to 0.38% for 500 attacks, with identical variability expressed through the CIs. Despite the increase in attack attempts, the privacy risk associated with the synthetic data remains effectively negligible, as reflected in the privacy risk value of 0.0 for both scales. The confidence intervals further support this observation, demonstrating tight bounds that underscore the robustness of the synthetic dataset against singling-out attacks. These findings are visually represented in both Figure 4.35 and Figure 4.36.

This evaluation underscores the AE model's effectiveness in generating synthetic datasets that uphold the privacy of individual records, making it a viable tool for sensitive data analysis where privacy preservation is paramount. These findings advocate for the utility of synthetic

data in research and applications requiring high standards of data privacy.

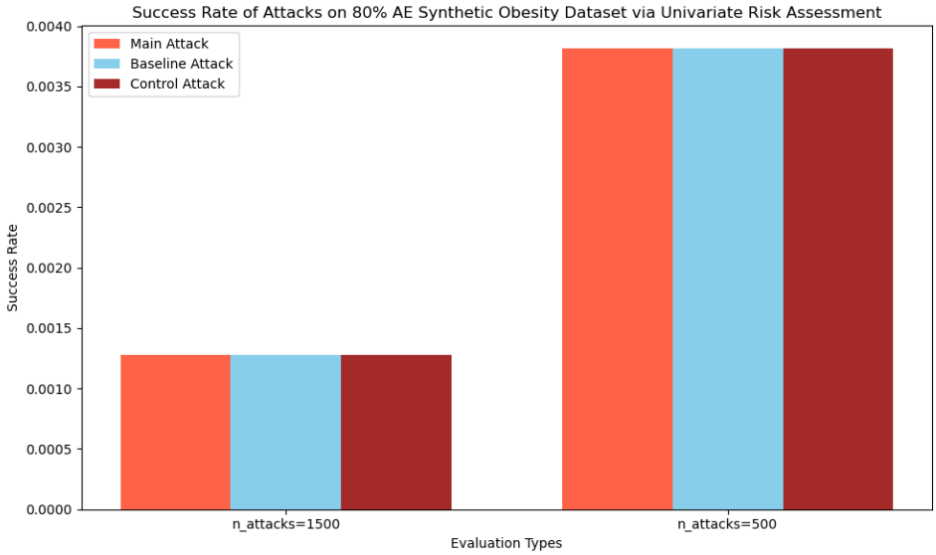


Figure 4.35: Comparison of Success Rates for Main, Baseline, and Control Attacks on AE Synthetic Obesity Data

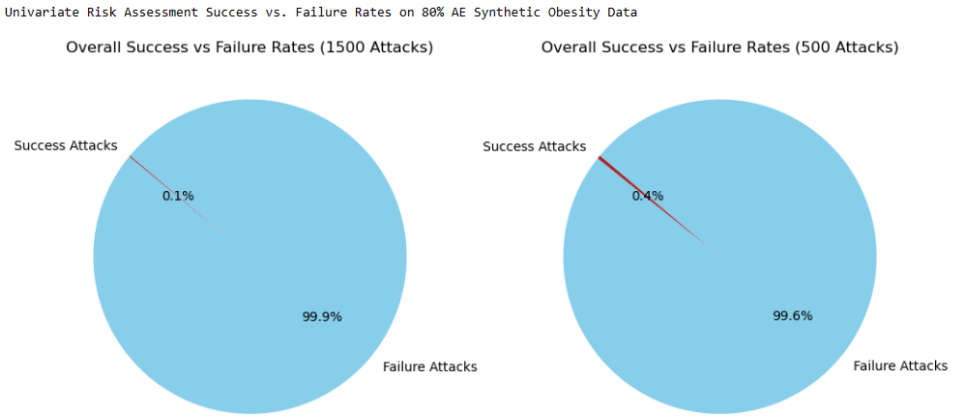


Figure 4.36: Overall Success and Failure Rates in Singling-Out Univariate Risk Assessment

4.3.2 Evaluation of Privacy Preservation through Singling-Out Multivariate Risk Assessment on AE Synthetic Obesity Data

Our exploration into the multivariate dimensions of privacy risks associated with the AE Synthetic Obesity Dataset, through an analytical lens of 1500 and 500 attacks, has unfolded a layered perspective on the dataset's privacy attributes. This analytical endeavor aims to unravel the intricacies of privacy preservation when multiple attributes are simultaneously considered for singling out individuals.
The analytical outcomes in Table 4.10 revealed through the risk assessment are as follows: A privacy risk of 0.0996 (with a confidence interval ranging from 0.0796 to 0.1196) for the 1500 attacks scenario, and a slightly lower risk of 0.0870 (with a confidence interval stretching from 0.0516 to 0.1225) for the 500 attacks scenario. Notably, the main attack success rate marginally decreases from 0.1336 to 0.1308 as we lower the attack count, reflecting a subtle yet discernible differential in the synthetic dataset's vulnerability.

Figure 4.37 below illustrates the success rates of attacks on the 80% AE Synthetic Obesity

Table 4.10: Singling-Out Multivariate Risk Assessment on 80% AE Synthetic Obesity Data

Evaluation Metric	**n_attacks=1500**	**n_attacks=500**
Main Attack Success Rate	0.1336 (Error: ±0.0172)	0.1308 (Error: ±0.0293)
Baseline Attack Success Rate	0.0132 (Error: ±0.0056)	0.0137 (Error: ±0.0095)
Control Attack Success Rate	0.0378 (Error: ±0.0096)	0.0480 (Error: ±0.0183)
Privacy Risk	0.0996 (CI: 0.0796 - 0.1196)	0.0870 (CI: 0.0516 - 0.1225)

Dataset via multivariate risk assessment, with evaluations conducted at 1500 and 500 attacks. Each bar represents a different type of attack: Main Attack, Baseline Attack, and Control Attack. The colors tomato, skyblue, and brown distinguish between these attacks, respectively. The x-axis categorizes the evaluation types based on the number of attacks, while the y-axis quantifies the success rate of each attack type. This visualization aids in comparing the effectiveness of the different attack types against the synthetic dataset, showcasing the dataset's resilience or vulnerability to each attack scenario.

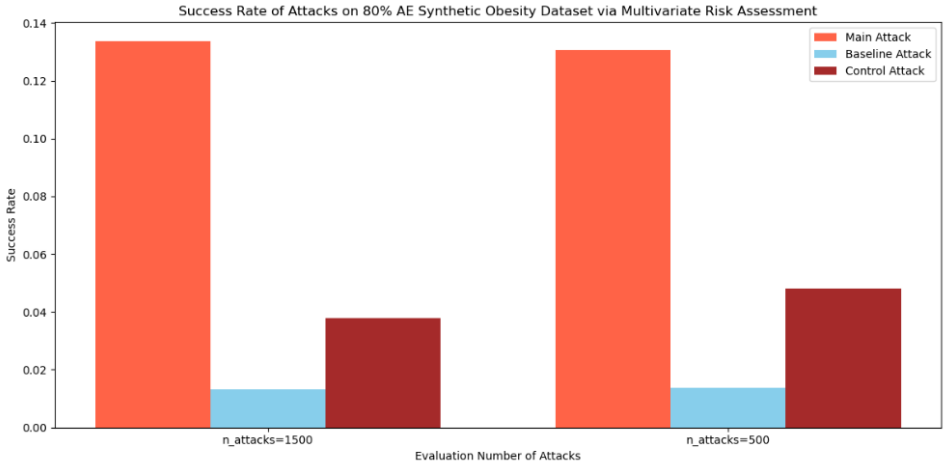


Figure 4.37: Success Rate of Attacks on 80% AE Synthetic Obesity Dataset via Multivariate Risk Assessment. The chart compares the success rates of Main, Baseline, and Control attacks at 1500 and 500 attacks, highlighting the synthetic dataset's security against singling out risks.

The graphical representation for the 1500 attacks scenario shed light on the 80% AE synthetic obesity dataset's resilience against singling-out attacks, manifesting through the main, baseline, and control attacks' success rates. The prominence of the main attack's success underscores the imperative for enhanced anonymization methods in synthetic data generation. The visualization for the 500 attacks scenario further underscores the criticality of robust data protection mechanisms. The increment in control attack success highlights the nuanced challenges faced in preserving privacy against more constrained adversarial attempts. These findings advocate for continuous advancements in the generation of synthetic datasets to uphold stringent privacy standards.

The pie charts in Figure 4.38 further simplify these observations into an overarching view of overall success versus failure rates for the two scenarios. For the 1500 attacks scenario, a more balanced distribution of success and failure rates highlights the AE Synthetic Data's resilience to singling-out attempts. Conversely, the 500 attacks scenario, while still predominantly showcasing overall failure, suggests a marginally higher success rate, indicating that the risk, though minimal, escalates slightly with fewer, more targeted attacks.

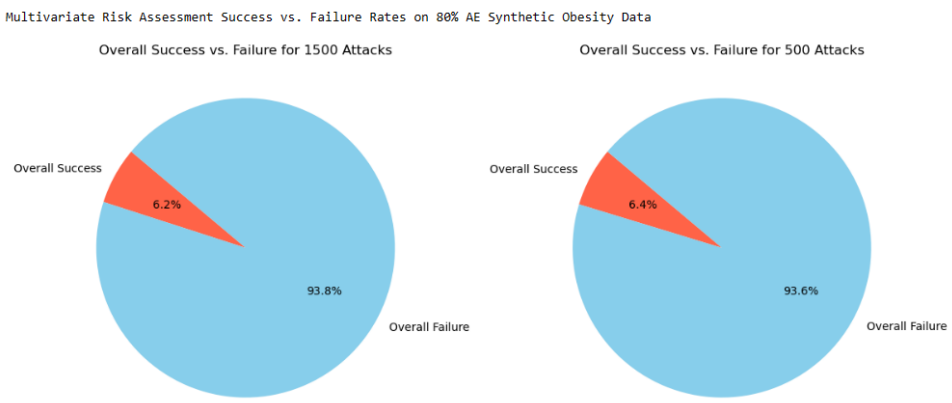


Figure 4.38: Success Rate of Attacks on 80% AE Synthetic Obesity Dataset via Multivariate Risk Assessment. The chart compares the success rates of Main, Baseline, and Control attacks at 1500 and 500 attacks, highlighting the synthetic dataset's security against singling out risks.

4.3.3 Evaluation of Privacy Preservation through Linkability Risk Assessment on AE Synthetic Obesity Data

Multivariate Risk Assessment

Evaluation Scenario	Linkability Risk	Privacy Risk	Confidence Interval (CI)	Success Rate (Main Attack)	Success Rate (Baseline Attack)	Success Rate (Control Attack)
n_neighbors=10	0.005278	Low	(0.0, 0.05297)	11.69%	5.60%	11.23%
n_neighbors=5	0.021955	Medium	(0.0, 0.05157)	6.07%	0.92%	3.96%

Figure 4.39: Tabular Representation of Linkability Risk Assessment.

Table 4.39 and subsequent visualizations provide a comprehensive overview of the linkability risk assessment conducted on 80% AE synthetic obesity disease data. The evaluation was carried out with two different scenarios based on the number of neighbors considered, nneighbors=10 and nneighbors=5, offering insights into how the granularity of neighbor consideration impacts the linkability and, by extension, the privacy risk. The analysis reveals a noteworthy distinction in linkability risk and success rates of attacks between the two scenarios. For nneighbors=10, the linkability risk is minimal, with a privacy risk labeled as low. This is indicative of the robust privacy-preserving capabilities of the AE synthetic data under conditions of broader neighbor consideration. The success rate of the main attack under this scenario stands at 11.69%, closely mirrored by the control attack's success rate of 11.23%, suggesting that the likelihood of accurately linking an individual's record between the original and synthetic datasets is marginally above random chance.

Conversely, the baseline attack's success rate of 5.60% underscores the synthetic data's efficacy in anonymizing individual identifiers. Shifting focus to the scenario with nneighbors=5, a nuanced increase in linkability risk is observed, pushing the privacy risk into a medium category. The reduction in neighbor consideration tightens the evaluation criteria, thereby slightly elevating the potential for successful linkage. Despite this uptick, the success rates across all attack types remain significantly low, with the main attack's success rate at 6.07%, further affirming the synthetic data's capability to safeguard against privacy infringements.

The bar charts in Figure 4.40 elucidate the comparative analysis of success rates across attack types for both scenarios, visually articulating the synthetic data's resilience against linkability attempts. The relatively flat distribution of success rates indicates a consistent defense mechanism embedded within the synthetic data generation process.

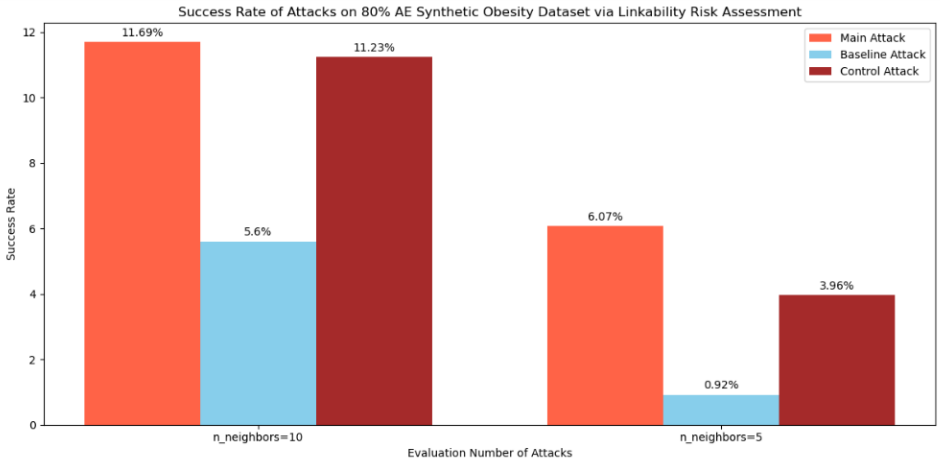


Figure 4.40: Distribution of Success Rates for Linkability Attacks. This bar chart compares the main, baseline, and control attack success rates for different neighbor settings, highlighting the effectiveness of linkability attacks under varying conditions.

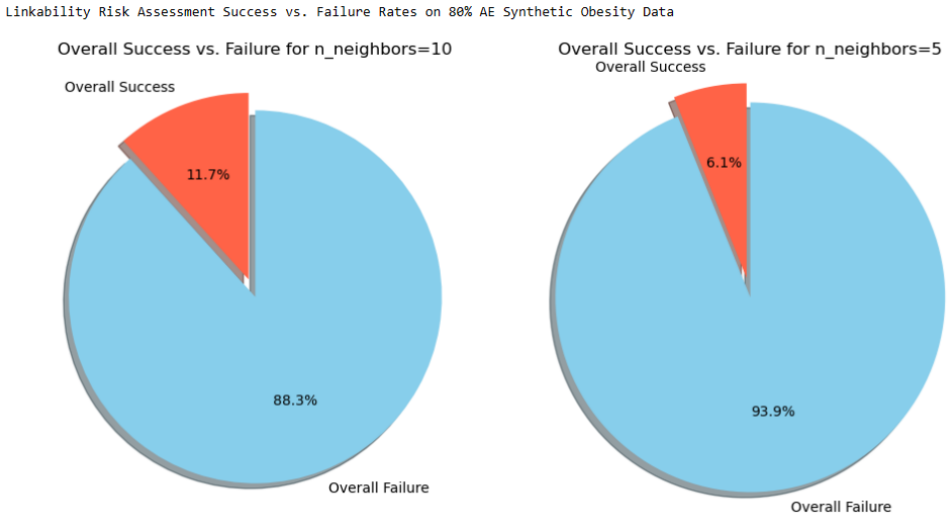


Figure 4.41: Overall Success vs. Failure Rate of Linkability Attacks. This pie-chart visualizes the proportion of successful linkability attacks against total attempts, providing a clear view of the attack's effectiveness in compromising privacy.

Pie charts in Figure 4.41 illustrating the overall success versus failure rates offer a stark visualization of the predominant failure rate in linking attempts. With over 88% failure rate for nneighbors=10 and an even more pronounced 93.93% for nneighbors=5, the graphical representation solidifies the synthetic data's role in maintaining individual privacy while providing a dataset viable for analytical pursuits. In summation, the conducted linkability risk assessment underscores the Autoencoder-generated synthetic obesity disease data's potential as a privacy-preserving tool. While the synthetic data showcases a favorable balance between utility and privacy, it also emphasizes the importance of context-specific evaluations to gauge the optimal parameters for synthetic data generation processes. This balance is crucial for leveraging synthetic data in sensitive domains like healthcare, where data utility and individual privacy must coexist harmoniously.

4.3.4 Evaluation of Privacy Preservation through Inference Risk Assessment Per-Column on AE Synthetic Obesity Data

Table 4.11 succinctly captures the inference risk across various attributes of the AE synthetic obesity dataset. Notably, it elucidates the privacy risk values alongside confidence intervals for each attribute, underscoring the dataset's susceptibility to inference attacks. Attributes with higher privacy risk values indicate areas where synthetic data might be more prone to allowing inference of sensitive information, thereby necessitating further scrutiny and potentially enhanced anonymization techniques.

Table 4.11: Inference Risk and Attack Success Rates

Attack Type	Success Rate	Failure Rate	Privacy Risk	CI-LBd	CI-UBd
Main Attack	88.77%	11.23%	0.108	0.0	0.225
Baseline Attack	16.38%	83.62%	0.235	0.0	0.490
Control Attack	79.40%	20.60%	0.162	0.0	0.364

Figure 4.42 delineates the privacy risk associated with each attribute within the synthetic dataset, offering a visual representation of vulnerability across the spectrum. Attributes with elevated privacy risk values, such as 'Height' and 'CH2O', highlight the pressing need for targeted improvements in data synthesis processes to curb the feasibility of inference attacks effectively. This visualization aids stakeholders in pinpointing specific data attributes that require additional protective measures to uphold individuals' privacy. Note: CI-LBd (confidence interval lower bound) and CI-UBd (confidence interval upper bound).

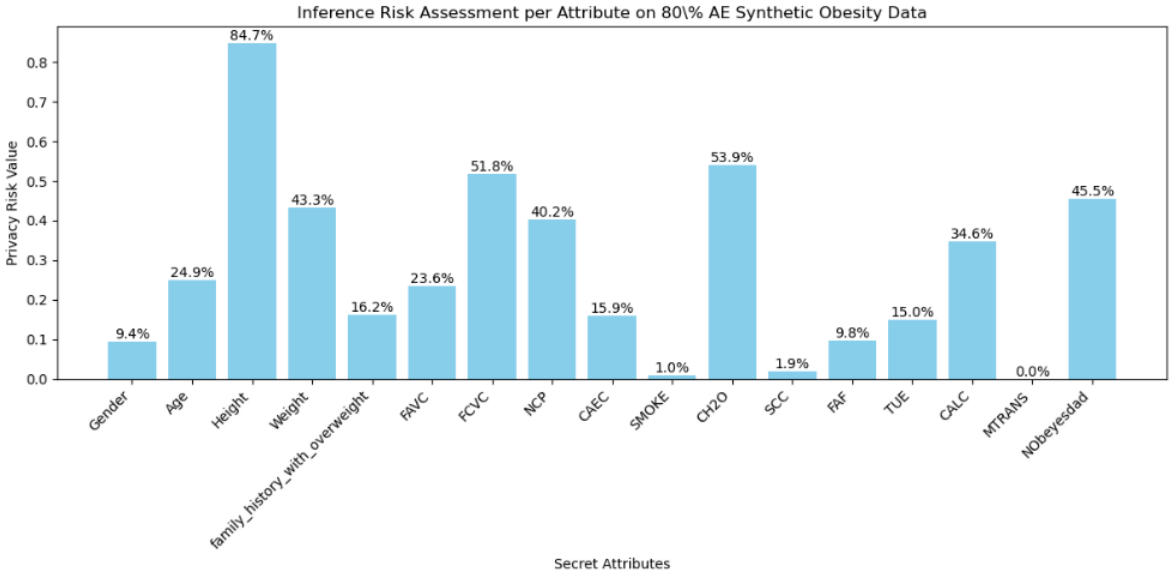


Figure 4.42: Overall Success vs. Failure Rate of Inference Attacks. This bar-chart visualizes the proportion of successful inference attacks against total attempts, providing a clear view of the attack's effectiveness in compromising privacy.

Figure 4.43 and Figure 4.44, depicting the overall success versus failure in safeguarding against inference attacks, manifest the synthetic dataset's resilience and areas of vulnerability in a stark, comprehensible manner. While the segments indicating success represent the dataset's effective anonymization in certain contexts, the failure segments draw attention to potential weaknesses, guiding efforts to fortify data privacy.

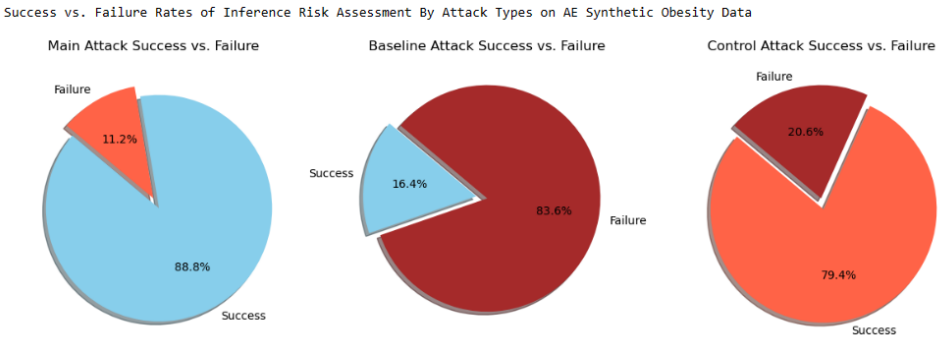


Figure 4.43: Overall Success vs. Failure Rate of Inference Attacks. This pie-chart visualizes the proportion of successful inference attacks against total attempts, providing a clear view of the attack's effectiveness in compromising privacy.

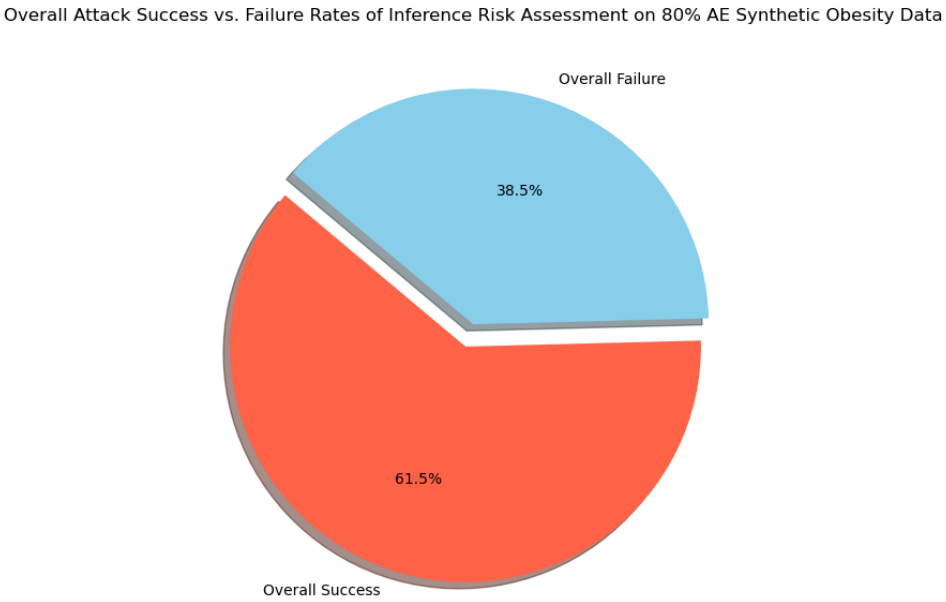


Figure 4.44: Overall Success vs. Failure Rate of Inference Attacks. This pie-chart visualizes the proportion of successful inference attacks against total attempts, providing a clear view of the attack's effectiveness in compromising privacy.

4.3.5 Evaluation of Privacy Preservation through Singling-Out Univariate Risk Assessment on VAE Synthetic Obesity Data

Table 4.12: Tabular Representation of Singling-Out Univariate Risk Assessment on 80% VAE Synthetic Obesity Data

Evaluation Metric	**n_attacks=1500**	**n_attacks=500**
Main Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Baseline Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Control Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Privacy Risk	0.0 (CI: 0.0, 0.0018); 0.0 (CI: 0.0, 0.0054)	

The privacy risk assessments conducted through singling out risk evaluations for a Variational Autoencoder (VAE) synthetic dataset of obesity disease data aim to quantify the risk of individual identification from the synthetic data. Two scenarios were evaluated: one with 1500

attacks and another with 500 attacks on the VAE synthetic obesity data. The results provide insights into the privacy preservation capabilities of the VAE model under these conditions. Key Findings:

- **Singling Out Risk:** The singling out risk, which measures the likelihood that an individual can be uniquely identified in the synthetic dataset, showed a value of 0.0 in both scenarios. This suggests that under the conditions tested, there is virtually no risk of singling out an individual, which is promising for privacy preservation.

- **Confidence Intervals:** The confidence intervals provide a range within which the true value of the privacy risk is expected to lie with a certain level of confidence. Narrower intervals indicate more precision in the risk estimation. For the 1500 attacks scenario, the upper bound of the confidence interval is 0.0018, and for the 500 attacks scenario, it is slightly higher at 0.0054. These low upper bounds further affirm the low privacy risk of the synthetic data.

- **Success Rates of Attacks:** The success rates of main, baseline, and control attacks are identical within each scenario, suggesting a consistent performance of the evaluation mechanism across different types of attacks. The success rates are extremely low (around 0.0013 for 1500 attacks and 0.0038 for 500 attacks), indicating a very low probability of successfully identifying individuals through attacks.

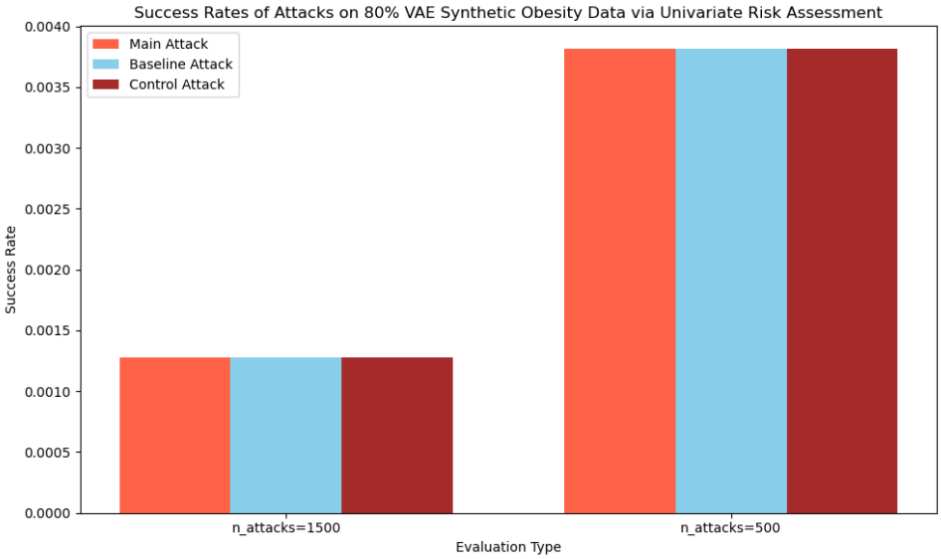


Figure 4.45: Comparison of Success Rates for Main, Baseline, and Control Attacks on VAE Synthetic Obesity Data

The bar chart in Figure 4.45 illustrates the number of attacks conducted in two scenarios: 1500 and 500 attacks on the VAE synthetic data. This visualization helps to compare the intensity of attacks directly, showing a more extensive evaluation in the first scenario.
The left Pie Chart (1500 Attacks) shows a very low success rate (0.13%) versus a failure rate of 99.87%, indicating that attempts to single out individuals in this scenario were largely unsuccessful, Similarly, the right Pie Chart (500 Attacks) demonstrates a success rate of 0.38% against a failure rate of 99.62%, again highlighting the effectiveness of the VAE in preserving privacy even with a reduced number of attacks, Figure 4.46.
The distinct pie charts for each attack scenario visually confirm the VAE model's robust privacy-preserving capabilities, with a clear majority of failure rates indicating the model's effectiveness against privacy attacks.

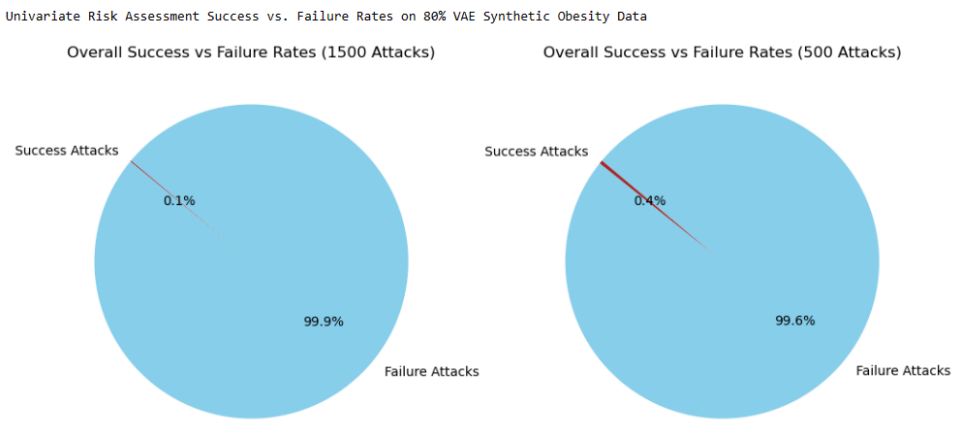


Figure 4.46: Overall Success and Failure Rates in Singling-Out Univariate Risk Assessment

4.3.6 Evaluation of Privacy Preservation through Singling-Out Multi-variate Risk Assessment on VAE Synthetic Obesity Data

Table 4.13: Singling-Out Multivariate Risk Assessment on 80% VAE Synthetic Obesity Data

Evaluation Metric	**n_attacks=1500**	**n_attacks=500**
Privacy Risk (%)	4.02 (CI: 2.43 - 5.60)	3.61 (CI: 0.92 - 6.30)
Main Attack Success Rate (%)	7.04 ± 1.29	6.73 ± 2.16
Baseline Attack Success Rate (%)	1.13 ± 0.52	1.57 ± 1.02
Control Attack Success Rate (%)	3.15 ± 0.88	3.24 ± 1.50

The multivariate risk assessment for the Variational Autoencoder (VAE) synthetic obesity symptoms data focuses on understanding the privacy risk when considering multiple attributes simultaneously. Two assessments were performed, one with 1500 attacks and the other with 500 attacks, analyzing four columns (attributes) in each case. Key results include.

For 1500 attacks, the privacy risk is assessed at 4.02%, with a confidence interval ranging from 2.43% to 5.60%. This indicates a moderate level of risk that an individual can be singled out based on the multivariate analysis. For 500 attacks, the privacy risk slightly decreases to 3.61%, with a wider confidence interval (0.92% to 6.30%). The wider interval suggests less certainty in the risk estimate due to fewer attacks. Success Rates of Attacks: In the 1500 attacks scenario, the main attack has a success rate of 7.04%, significantly higher than the baseline attack (1.13%) and the control attack (3.15%). This suggests that the synthetic data, under certain conditions, can be prone to identification risks, albeit at a relatively low probability. In the 500 attacks scenario, the main attack's success rate is slightly lower at 6.73%, with baseline and control attacks having success rates of 1.57% and 3.24% respectively. Despite fewer attacks, the risk pattern remains consistent with the 1500 attacks scenario.

The multivariate risk assessments indicate a consistent but moderate privacy risk across both scenarios, suggesting that while the synthetic data generated by the VAE model does maintain individual privacy to a large extent, there remains a non-negligible risk of re-identification in certain conditions. The success rates of the main attacks, higher than those of baseline and control attacks, highlight the effectiveness of targeted attacks over random guessing or control conditions, reinforcing the need for careful consideration of privacy risks when using synthetic data in sensitive contexts.

The vertical bar chart in Figure 4.47 accurately reflects singling out risk assessment results for the VAE synthetic obesity data under multivariate conditions. Each set of bars corresponds to

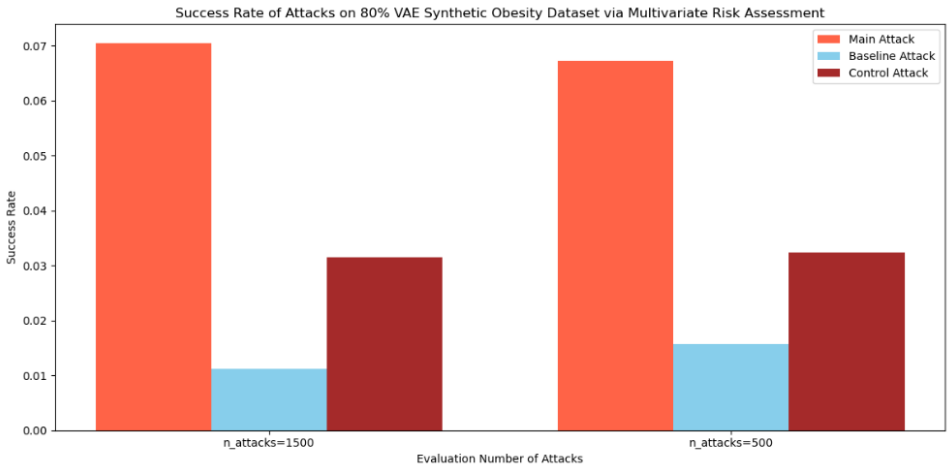


Figure 4.47: Comparison of Success Rates for Main, Baseline, and Control Attacks on VAE Synthetic Obesity Data

a specific type of attack (Main Attack, Baseline Attack, Control Attack) for different numbers of attacks (1500 and 500). Main Attack Success Rates show that the success rates of the main attack slightly decrease when the number of attacks is reduced from 1500 to 500, suggesting a consistent level of vulnerability across different attack volumes. Baseline vs. Control Attack Success Rates reveal that both baseline and control attacks show lower success rates compared to the main attacks, indicating that targeted approaches are more effective in identifying individuals within the synthetic dataset. Comparative Analysis indicates that the relatively close success rates between baseline and control attacks across both attack volumes underscore a level of risk inherent in the dataset, regardless of the attack sophistication. However, the main attack's higher success rate highlights specific vulnerabilities that could potentially be exploited.

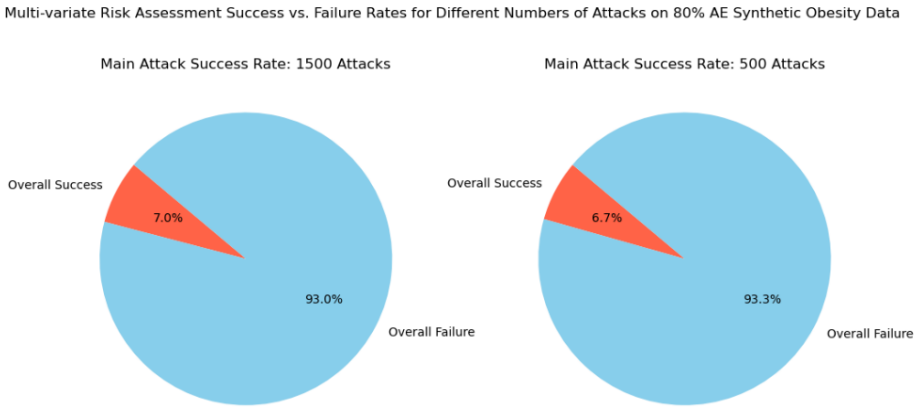


Figure 4.48: Overall Success and Failure Rates in Singling-Out Multivariate Risk Assessment

Main Attack Success Rate for 1500 Attacks in Figure 4.48 shows the pie chart illustrates that the success rate of the main attack in the 1500 attacks scenario is 7.04%, indicating a relatively low but notable risk of identifying individuals within the synthetic dataset. Main Attack Success Rate for 500 Attacks similarly shows that for 500 attacks, the success rate slightly decreases to 6.73%. Despite fewer attacks, this slight reduction in success rate suggests a consistent underlying risk pattern across different levels of attack intensity.

4.3.7 Evaluation of Privacy Preservation through Linkability Risk Assessment on VAE Synthetic Obesity Data

The linkability risk assessment conducted on 80% VAE synthetic obesity data explores the risk associated with correctly linking records between the original and synthetic datasets, focusing on scenarios with different numbers of neighbors (nneighbors=10 and nneighbors=5) in the analysis.

Privacy Risk analysis reveals that when nneighbors=10, the privacy risk is assessed at 0.0%, with a confidence interval ranging up to 0.0081%. This extremely low risk suggests that it's highly unlikely for an attacker to link records between the datasets based on the given attributes. Similarly, when nneighbors=5, the privacy risk remains 0.0%, with a confidence interval extending up to 0.0081%, echoing the findings from the nneighbors=10 scenario.

Success Rates of Attacks discussion highlights that with nneighbors=10, the main attack success rate is 6.07%, which is slightly lower than the control attack rate of 8.88%, indicating that random guessing might sometimes be more effective than the main attack. With nneighbors=5, the success rates for both main and baseline attacks drop significantly to 1.39%, while the control attack rate is 2.32%. This reduction further underscores the difficulty in successfully linking records as the criteria for linking (nneighbors) becomes stricter.

The consistent low privacy risk across both scenarios indicates a strong privacy-preserving capability of the VAE synthetic dataset against linkability attacks, even when the attack model is adjusted. The relatively low success rates of the main and baseline attacks, particularly with nneighbors=5, further affirm the synthetic dataset's resilience to linkability risks, highlighting its potential for use in scenarios where privacy is a critical concern.

n_neighbors	Privacy Risk (%)	Confidence Interval	Main Attack Success Rate (%)	Baseline Attack Success Rate (%)	Control Attack Success Rate (%)
10	0.0	(0.0, 0.0081)	6.07 ± 2.23	6.31 ± 2.27	8.88 ± 2.67
5	0.0	(0.0, 0.0081)	1.39 ± 1.02	1.39 ± 1.02	2.32 ± 1.36

Figure 4.49: Overall Success and Failure Rates in Linkability Risk Assessment

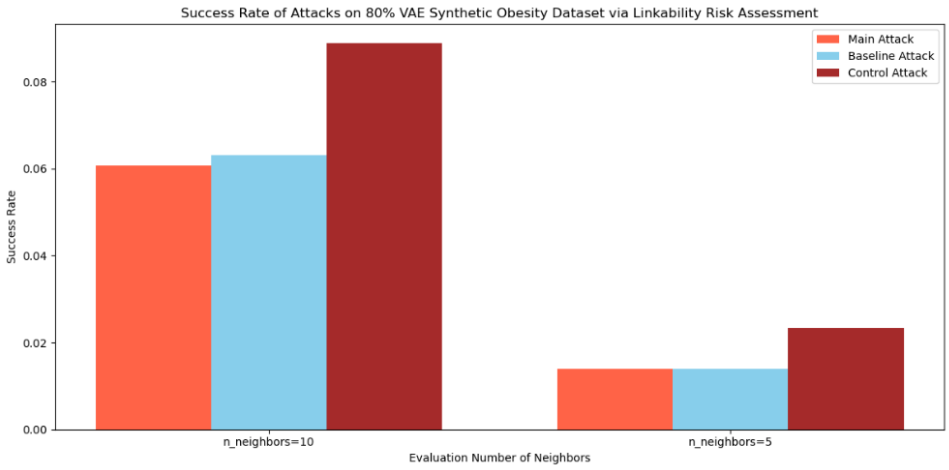


Figure 4.50: Comparison of Success and Failure Rates of Different Neighbors in Linkability Risk Assessment

In Figure 4.50 the vertical bar chart analysis delves into Main Attack Success Rates, which

are illustrated for both scenarios, accentuating a decrease in success rate when the number of neighbors is reduced from 10 to 5. This observation suggests that as the criteria for linking becomes stricter, the ability to successfully link records diminishes. Baseline vs. Control Attack Success Rates are likewise presented, illustrating how both types of attacks compare against the main attack. Notably, for nneighbors=10, the control attack success rate surpasses both the main and baseline attacks, suggesting an inherent level of risk present within the control conditions.

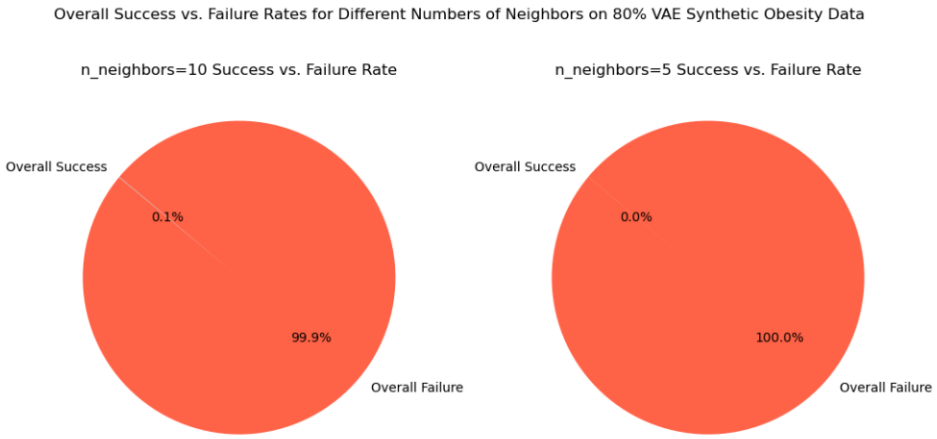


Figure 4.51: Overall Success and Failure Rates in Linkability Risk Assessment

Pie Charts Analysis in Figure 4.51 explores Success vs. Failure Rates for nneighbors=10 and 5, providing a vivid visual representation of the main attack's effectiveness under these two distinct settings. With nneighbors set to 10, the success rate is approximately 6.07%, indicating a modest but significant risk of successful linkage. Conversely, when the neighbor count is reduced to 5, there's a slight decrease in success rate, further emphasizing the protective effect of stricter linking criteria.

These analyses underscore the nuanced balance between utility and privacy in synthetic datasets. While the VAE model demonstrates a strong capability to preserve individual privacy against linkability attacks, the assessments reveal areas where privacy risks persist, emphasizing the importance of careful consideration in deploying synthetic data, especially in sensitive applications.

4.3.8 Evaluation of Privacy Preservation through Inference Risk Assessment Per-Column on VAE Synthetic Obesity Data

The inference risk assessment on 80% VAE synthetic obesity data offers an in-depth analysis of the potential privacy risks associated with various attributes, determined through the ability to infer sensitive information accurately. By using the smallest dataset size as a benchmark for the number of attacks attempted, this assessment provides a realistic perspective on potential data inference vulnerabilities. Key Findings from Inference Risk Assessment.

Table 4.14: Inference Risk and Attack Success Rates on VAE Synthetic Obesity Data

Attack Type	Success Rate	Failure Rate	Privacy Risk	CI LBd	CI UBd
Main Attack	59.96%	40.04%	0.599	0.551	0.646
Baseline Attack	13.10%	86.90%	0.131	0.099	0.163
Control Attack	59.25%	40.75%	0.592	0.545	0.638

Table 4.14 presents the results of an inference risk assessment performed on 80% VAE synthetic obesity data, highlighting the effectiveness of main, baseline, and control attacks in inferring

sensitive information. Success rates indicate the proportion of attacks that accurately predicted private information, while failure rates show the proportion that did not. Privacy risk quantifies the likelihood of an individual's data being correctly inferred, with confidence intervals (CI Lower Bound and CI Upper Bound) providing the statistical range of this risk estimate. These metrics together offer insight into the data's vulnerability to privacy breaches and the relative security afforded by synthetic data generation techniques.

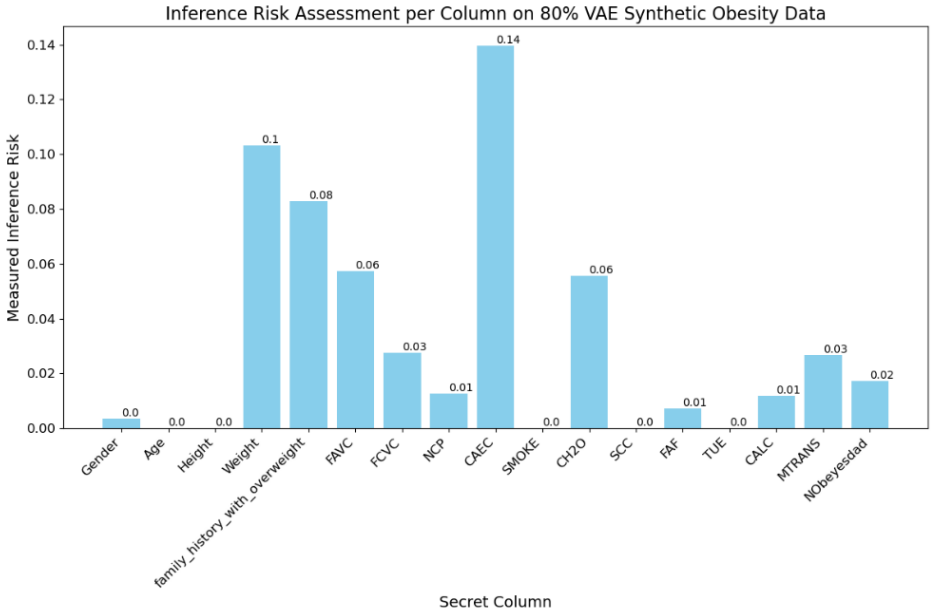


Figure 4.52: Inference Risk Assessment Per Column of VAE Synthetic Obesity Data

The assessment in Figure 4.52 uncovers attribute-specific privacy risks, notably identifying CAEC, Weight, (FHOW), FAVC, CH2O, FCVC, and MTRANS as the attributes bearing the highest privacy risks, marked at 14.00%, 10.00%, 8.00%, 6.00%, 6.00% and 3.00% respectively. These results underscore the particular vulnerabilities within the synthetic dataset, emphasizing the need for targeted privacy-preserving measures to mitigate risks associated with these attributes.

The main attack demonstrated a high success rate of 60.00%, indicating a significant capability to accurately infer data. Conversely, the baseline attack yielded a much lower success rate of 13.20%, whereas the success rate of the control attack was nearly as high as that of the main attack, standing at 59.30%. This distribution of success rates across different types of attacks highlights the varying levels of threat they pose to the privacy of individuals represented in the synthetic dataset.

The pie charts in Figure 4.53 provide a visual representation of the success versus failure rates for the different types of attacks conducted on the 80% VAE synthetic obesity data, including an overall assessment that combines all attack types. Analysis from Figure 4.53 highlight several key points:

The Overall Success vs. Failure Rate shows that the combined success rate from all types of attacks reflects a significant ability to infer sensitive information from the synthetic dataset, underscoring the need for enhanced privacy-preserving measures in VAE models.

The Main Attack Success Rate is particularly high, signifying a substantial risk and indicating that targeted attacks are especially effective in exploiting the synthetic dataset's vulnerabilities.

The success rate of the Baseline Attack is notably lower, suggesting that more generic, less

targeted attempts at inference are less likely to succeed, though they are not entirely without risk.

The Control Attack, with a success rate close to that of the Main Attack, suggests that even in controlled conditions, where specific strategies might not be employed, there's still a significant risk of successful inferences being made. This comprehensive view aids in understanding the various facets of privacy risks associated with synthetic data, guiding future strategies for data protection.

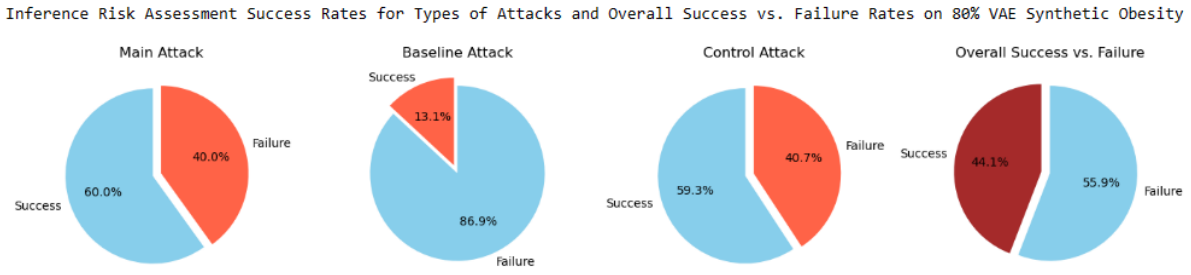


Figure 4.53: Overall Success and Failure Rates in Inference Risk Assessment

4.3.9 Comparative Analysis of Privacy Risk Assessment Between AE and VAE Synthetic Obesity Datasets

This analysis compares the privacy risk assessments of synthetic obesity datasets generated by Autoencoder (AE) and Variational Autoencoder (VAE) models, focusing on their effectiveness in preserving privacy against various types of privacy attacks.

Singling-Out Univariate Risk Assessment

Results from the AE dataset indicate extremely low risk levels in both tested scenarios (1500 and 500 attacks), with success rates barely reaching 0.0038%, suggesting effective privacy preservation. Refer to Table 4.9 and Figures 4.35 and 4.36 for a detailed view.

Similarly, the VAE dataset shows almost identical success rates and risk levels in corresponding scenarios, indicating a comparable level of privacy preservation effectiveness. See Table 4.12 and Figures 4.45 and 4.46 for more details.

Multivariate Risk Assessment

The multivariate risk assessment on the AE dataset reveals moderate privacy risks, with success rates for main attacks slightly exceeding those for baseline and control attacks, suggesting potential vulnerabilities when multiple attributes are considered. Detailed results can be found in Table 4.10.

The VAE dataset, however, presents a slightly increased privacy risk in multivariate conditions compared to AE, with main attack success rates slightly higher, indicating a marginal decrease in privacy preservation capability under complex attack scenarios. Refer to Table 4.13 for specifics.

Linkability Risk Assessment

The AE synthetic dataset demonstrates strong defenses against linkability attacks, with a low probability of successfully linking individual records between the original and synthetic datasets. The main attack success rates are relatively low, reflecting robust anonymization techniques. For detailed metrics, refer to Figure 4.38 which illustrates the success rates

for various types of attacks and Table 4.39 which provides a tabular overview of the risk assessments.

In the VAE synthetic dataset, the linkability risk assessments yield similar outcomes, with low success rates for main, baseline, and control attacks, indicating a strong capability to prevent record linkage. This confirms the effectiveness of the VAE in maintaining privacy against linkability threats. The details of these assessments are presented in Table 4.49 and Figure 4.50, showing a comparison of attack success rates under different settings.

Inference Risk Assessment

Inference risks in the AE dataset were significant, particularly under main attacks, highlighting vulnerabilities in inferring sensitive information. Consult Table 4.11 for a comprehensive overview.

VAE exhibited a similar pattern, with main attacks again showing high success rates. However, the control and baseline attacks in VAE performed comparably to those in AE, suggesting that both models share similar risks in terms of inference capabilities. See Table 4.14 for detailed results.

Both AE and VAE models exhibit strong privacy-preserving properties under basic attack scenarios but show vulnerabilities under more complex multivariate and inference attacks. The consistency in their performance suggests that both models derive from similar underlying principles in handling privacy, albeit with slight differences in handling complex scenarios. In the same manner, both AE and VAE synthetic datasets exhibit comparable and robust privacy preservation capabilities in the context of linkability risk assessments. Although both datasets show low success rates in linking individual records, subtle differences in the configurations of attack types suggest that while both models effectively anonymize data, they might do so with varying efficiencies under different attack scenarios. These findings are crucial for stakeholders considering the use of synthetic datasets for analyses where linkability poses a significant risk.

In a nutshell, the analysis suggests that both AE and VAE synthetic datasets maintain high standards of privacy preservation against univariate attacks but may require additional safeguards against multivariate and inference attacks. Future work should focus on enhancing the models' robustness against more sophisticated attacks to ensure comprehensive privacy preservation.

4.4 Data Utility of Cardiovascular Disease Data

4.4.1 Comparative Analysis of Original and AE Synthetic Cardiovascular Disease Data: A Multi-Faceted Evaluation Using AE Model

In this section, we performed Chi-Squared tests to evaluate the independence between observed distributions of categorical features in the original cardiovascular training data and the AE synthetic dataset. The Chi-Squared statistic measures how expectations compare to actual observed data. A higher statistic suggests greater divergence from expected independence.

The results, summarized in Table 4.16, reveal varying degrees of dependency between the datasets' categorical features. Notably, 'Glucose' and 'Cholesterol' show a lower P-value (0.0331 and 0.0329 respectively), indicating a statistically significant difference in their distributions compared to others like 'Smoke' and 'Active', which exhibited high P-values, suggesting a closer match to expected distributions under the null hypothesis of independence. These findings are crucial for assessing the quality of synthetic data in replicating the statistical

characteristics of original data, impacting the reliability of synthetic datasets for further cardiovascular studies.

Table 4.15: Chi-Squared Test Results for Categorical Features

Feature	Chi-Squared Stat	P-value	Degrees of Freedom
Gender	7008.93	0.4778	7003
Cholesterol	23634.24	0.0329	23236
Glucose	38220.23	0.0331	37714
Smoke	43051.74	0.6531	43168
Alcohol	44248.13	0.3762	44155
Active	10712.33	0.9960	11103
Cardio	0.443	0.5056	1

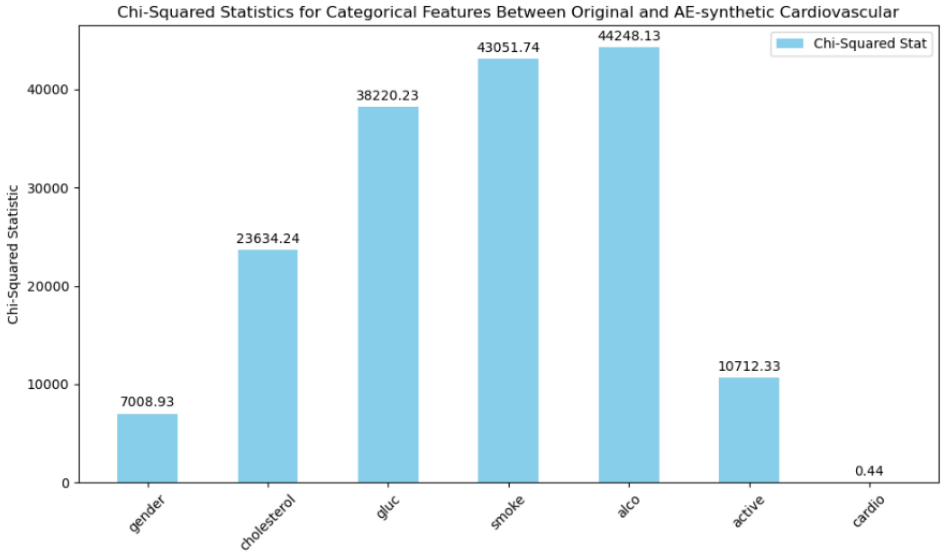


Figure 4.54: Chi-Squared Statistics for Categorical Features

The bar graph 4.54 illustrates the Chi-Squared statistics for various categorical features within the dataset. Each bar represents a different feature, such as gender, cholesterol, gluc, smoke, alcohol consumption (alco), physical activity (active), and cardiovascular condition (cardio). The height of each bar corresponds to the Chi-Squared statistic value for that feature, providing a visual comparison of the extent to which the distributions in the original and synthetic datasets differ for each category. Notably, features like smoke, alco, and gluc show higher Chi-Squared values, suggesting greater discrepancies between the datasets in these categories. In contrast, the cardio feature has a significantly lower Chi-Squared statistic, indicating minimal difference in the distribution of this category between the two datasets. This graphical representation aids in easily identifying which features might need further alignment or adjustment in the synthetic data generation process to more closely mimic the original data's categorical distributions.

Comparative Analysis of Performance Metrics Between Original and AE Synthetic Cardiovascular Data

Table 4.16 and subsequent graphical analyses provide a detailed comparison of various performance metrics across original and autoencoder-generated synthetic cardiovascular data. The metrics analyzed include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Accuracy for categorical variables. Notably, the synthetic data exhibits lower error rates in continuous variables, indicating high fidelity in the synthetic

Table 4.16: Comparison of Error Metrics and Accuracy Between Original and AE Synthetic Cardiovascular Data

Feature	MSE	RMSE	MAE	Accuracy (%)
Age	0.0071	0.0843	0.0660	-
Gender	-	-	-	65.00
Height	0.3861	0.6214	0.4669	-
Weight	0.0777	0.2788	0.2242	-
Ap_hi	0.9272	0.9629	0.0923	-
Ap_lo	0.6072	0.7792	0.2582	-
Cholesterol	-	-	-	86.98
Gluc	-	-	-	92.16
Smoke	-	-	-	92.51
Alco	-	-	-	94.52
Active	-	-	-	92.64

generation process. The accuracy values for categorical variables such as Gender, Cholesterol, and Alcohol Consumption are reasonably high, demonstrating that the synthetic dataset preserves essential categorical characteristics effectively.

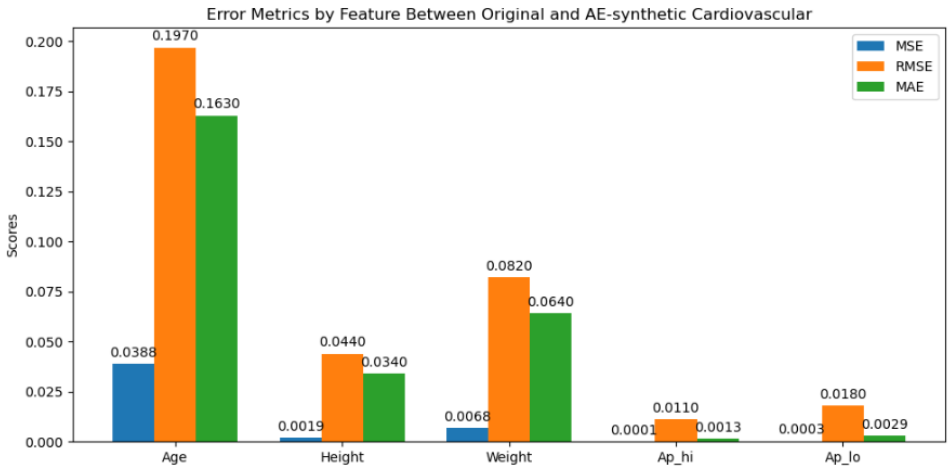


Figure 4.55: Comparative Analysis of Error Metrics and Accuracy between Original and VAE Synthetic Cardio Data

The bar graph 4.55 visually compares the error metrics MSE, RMSE, and MAE for different features related to cardiovascular data. Each bar represents a different metric for a specific feature, providing a clear visual indicator of the data's quality and consistency. By using different colors for each metric and aligning them with their respective feature, the graph offers an intuitive understanding of where discrepancies or issues might exist. This visualization aids in quickly identifying features with higher errors and assessing the synthetic data's fidelity compared to the original.

Correlation Matrix Comparison Between Original and AE-Synthetic Cardiovascular Disease Data

The correlation matrices for both the 80% original and the 80% AE synthetic cardiovascular disease datasets provide a clear view of how each attribute relates to others within the same dataset. Correlation values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation.

- **Original Cardiovascular Disease Data: Strong Correlations**, notably, the original

dataset shows strong correlations between cholesterol and glucose **(0.46)**, which is expected as both can be indicators of metabolic health. **Gender and Height/Smoke**, the gender shows strong correlations with height **(0.50)** and smoke **(0.34)**, suggesting that gender might influence body height and smoking habits in this dataset. **Weight Correlations**, weight shows moderate positive correlations with several other health indicators like height **(0.29), cholesterol (0.13), and glucose (0.11)**. Information can be found in Figure 4.56.

- **AE-Synthetic Cardiovascular Disease Data: Heightened Correlations in Synthetic Data**, the synthetic data displays a notably stronger correlation between gender and height **(0.77)** compared to the original data. This could suggest that the synthetic generation process might be exaggerating relationships found in the original data or that the model used for synthesis captured these relationships with high sensitivity. **Cholesterol and Gluc**, similar to the original data, cholesterol, and gluc show a significant correlation **(0.46)**, indicating that the synthetic model effectively captures relationships relevant to metabolic health. **Alco and Smoke**, both datasets show similar patterns where lifestyle choices like alcohol consumption and smoking are correlated **(0.35 in the original and 0.35 in synthetic)**. Information can be found in Figure 4.57.

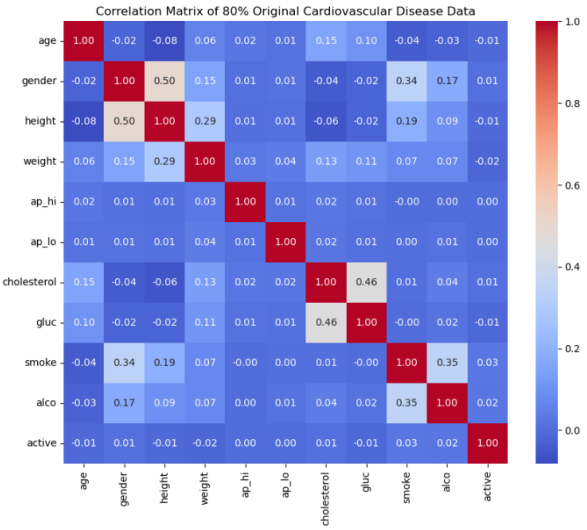


Figure 4.56: Original Cardiovascular Data

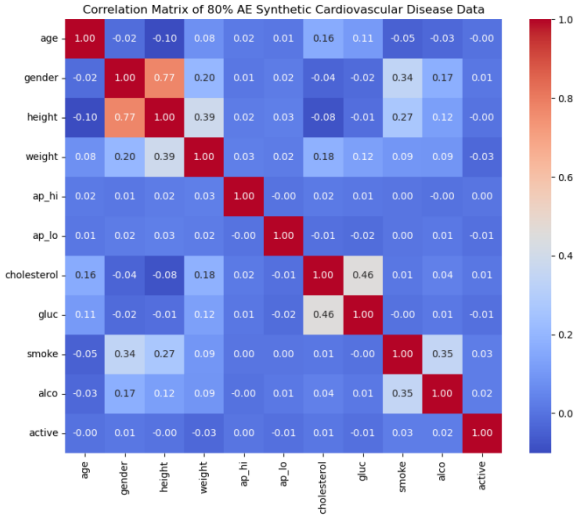


Figure 4.57: AE Synthetic Cardiovascular Data

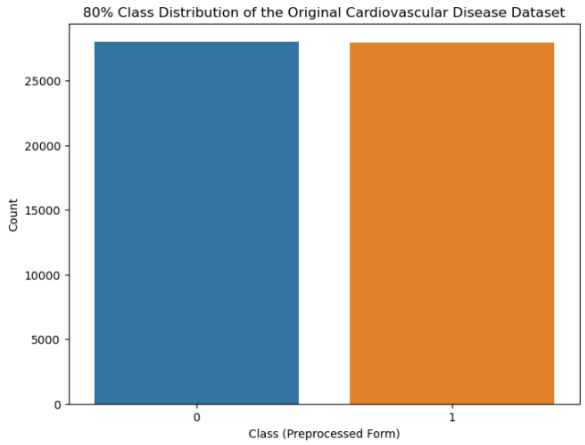


Figure 4.58: Class Distribution Original

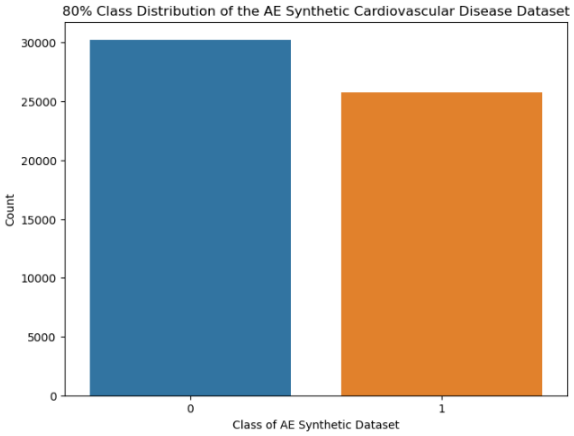


Figure 4.59: Class Distribution AE Synthetic

The graph in Figure 4.60 visualizes the strength of relationships between various health

indicators in both the original and synthetic datasets. It highlights where synthetic data may overestimate or accurately represent these relationships. Certain key Observations were made. Increased correlation in synthetic data for gender and height suggests potential overfitting or biases introduced during the data generation process. Similar patterns in health-related correlations (cholesterol with gluc) across datasets indicate good model performance in capturing relevant health data interactions. These visuals and interpretations provide clear insights into the nature of the synthetic data relative to the original, highlighting both strengths in replicating important relationships and areas where the synthesis may not perfectly align with real-world data dynamics.

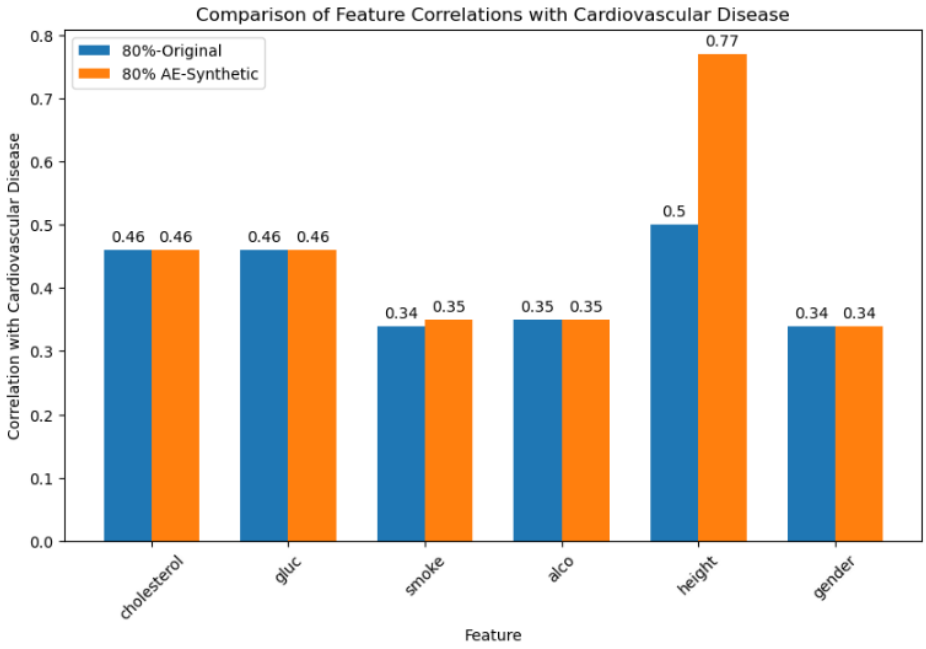


Figure 4.60: Correlation Coefficients Bars Comparison Between Original and AE-Synthetic Cardiovascular Disease Data

The scatter plot in Figure 4.61 compares the correlation coefficients between various features and the presence of cardiovascular disease (cardio) across the original and Autoencoder (AE) synthetic datasets. Each point represents a feature, plotted based on its correlation coefficient with cardiovascular disease in the original dataset (x-axis) against the AE synthetic dataset (y-axis). The dashed red line illustrates where the correlation coefficients would lie if they were identical in both datasets. **Close Alignment:** Most features lie near the red line, suggesting that the synthetic dataset preserves the correlation structure of the original dataset well. **Feature Deviations:** Points off the line indicate deviations, where the synthetic dataset's correlations either underestimate or overestimate the original data's values. For example, height and weight and aplo in the synthetic dataset show a higher correlation with cardiovascular disease than in the original dataset, potentially indicating overfitting or excessive modeling of these features. **Cholesterol, age, alco, gender, active, and gluc:** Both show very close correlation values between the datasets, highlighting the VAE model's capability to effectively mimic complex biochemical feature interactions. **Insight on Data Quality:** This graphical representation helps assess the fidelity of the synthetic dataset in replicating important statistical characteristics of the original data, which is crucial for ensuring that the synthetic data is useful for downstream tasks like predictive modeling and risk assessment. Figure 4.61 The classifiers in Figure 4.62 exhibit varying degrees of performance on both datasets, with most showing significantly higher mean cross-validation (CV) accuracy on the AE-Synthetic dataset compared to the Original dataset. This suggests that the AE-Synthetic dataset might be providing a more consistent or cleaner representation of the underlying

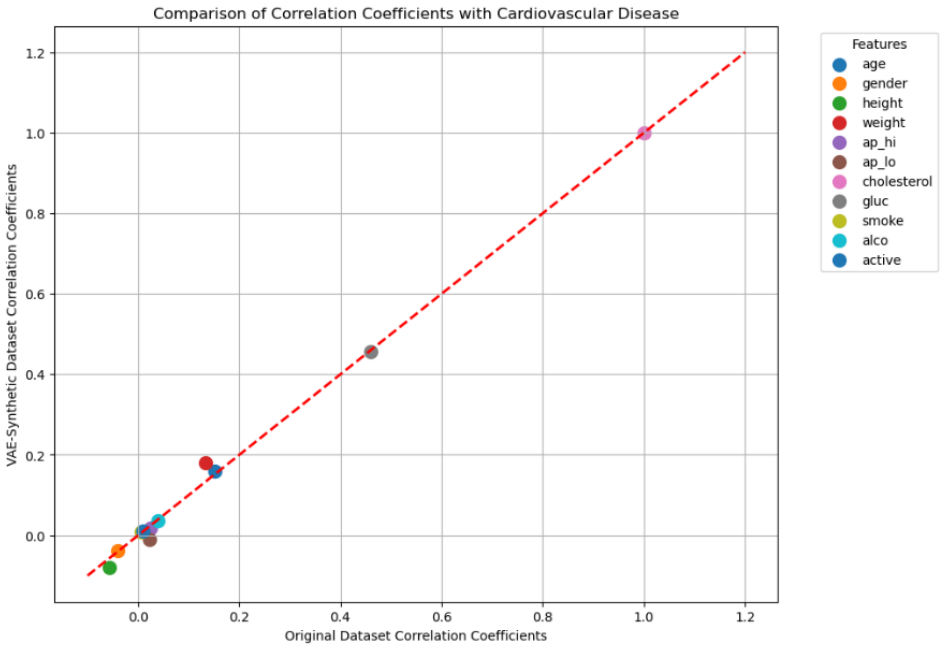


Figure 4.61: Correlation Coefficients Scatter Plots Comparison Between Original and AE-Synthetic Cardiovascular Disease Data

patterns associated with cardiovascular disease, potentially due to the data generation process smoothing over noise present in the real data. This graphical representation serves to elucidate the performance disparities and similarities across various machine learning models when evaluated on two distinct datasets.

Notably, classifiers like **RandomForest, XGBClassifier, LGBM, KNeighborsClassifier, and DecisionTreeClassifier** exhibit pronounced performance on the AE-Synthetic dataset with accuracies of **0.9418, 0.9355, 0.9224, 0.9218, and 0.9133** respectively, outperforming their counterparts on the original dataset, where they achieved **0.7143 and 0.7293** respectively. This significant discrepancy underscores the potential of synthetic datasets in enhancing classifier performance. Figure 4.62

Conversely, models such as the **DecisionTreeClassifier and KNeighborsClassifier** demonstrate a substantial variance in performance between the two datasets. For instance, the **DecisionTreeClassifier's** accuracy jumps from **0.6349** on the original dataset to **0.9133** on the AE-Synthetic dataset, highlighting how synthetic data can dramatically influence the model's ability to generalize. Figure 4.62

On the other hand, **LogisticRegression and SVC** exhibit more consistent performance across both datasets, with accuracies hovering around **0.71 to 0.73** for the original dataset and **0.81 to 0.85** for the synthetic dataset, indicating a moderate enhancement when using AE-Synthetic data. Figure 4.62

This comparison not only showcases the variance in model performances between the original and synthetic datasets but also emphasizes the importance of dataset quality and relevance in machine learning applications. It suggests that synthetic data, particularly those generated through sophisticated methods like autoencoders, can be a valuable asset in training more accurate and robust models, especially in scenarios where original data might be limited or biased.

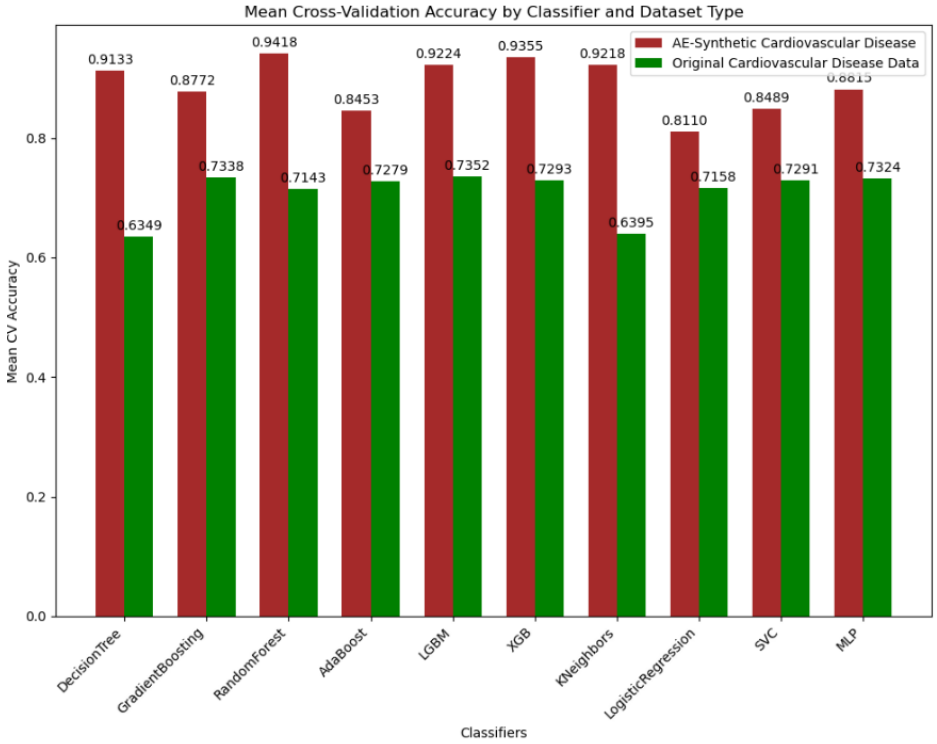


Figure 4.62: Mean Cross-Validation Accuracy By Classifier Between 80% Original and AE-Synthetic Cardiovascular Disease Data - Back downward to Table 4.41

Classification Reports By Classifier Between Original and AE-Synthetic Cardiovascular Disease Data

In the comprehensive assessment of classifier performance through classification reports, we observed varied outcomes across different classifiers when tested on both original and AE-synthetic datasets of cardiovascular disease. For the Random Forest classifier, precision and recall demonstrated close alignment on the original dataset, yielding a balanced f1-score of 0.72 for both classes. On the AE-synthetic dataset, this classifier managed a slightly better performance, achieving an accuracy of 0.73 with a precision of 0.79 for class 1, though the recall for class 1 dropped to 0.62, indicating a preference for precision over recall in this instance. Figure 4.63.

Gradient Boosting classifiers displayed a similar trend, where performance on the original dataset resulted in a slightly higher accuracy of 0.74, compared to 0.72 on the AE-synthetic dataset. Notably, the recall for class 0 on the AE-synthetic dataset was higher (0.84), suggesting better identification of true positives for this class compared to the original dataset. Figure 4.63.

For Logistic Regression, the accuracy was consistent across both datasets, achieving 0.73 on the original and a slightly lower 0.67 on the AE-synthetic. The MLP Classifier showed a robust performance with the highest accuracy observed of 0.75 on the original dataset and 0.71 on the AE-synthetic dataset. It managed better precision for class 1 on the synthetic dataset but at the cost of a lower recall rate. Figure 4.63.

Interestingly, the SVC and KNN classifiers showed modest accuracies with SVC achieving 0.74 and 0.69 and KNN at 0.67 and 0.65 on the original and AE-synthetic datasets respectively. The DCT (Decision Tree Classifier) presented the least favorable outcomes, with accuracy pegged at 0.63 across both datasets, highlighting its struggle with balancing precision and recall across class labels. Figure 4.63.

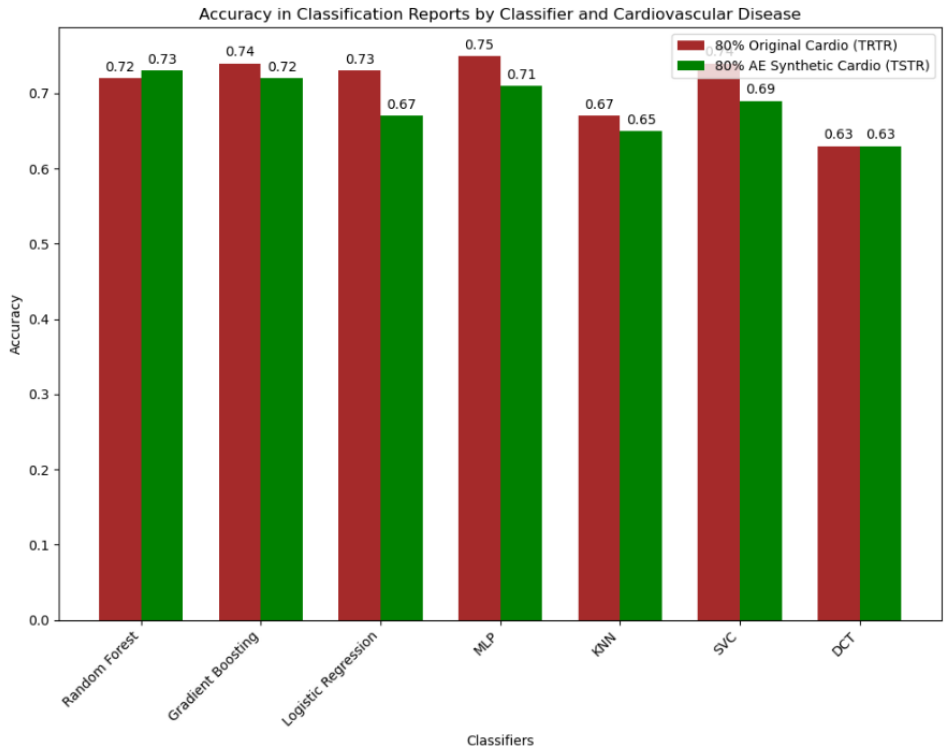


Figure 4.63: Classification Reports By Classifier Between Original and AE-Synthetic Cardiovascular Disease Data

Area Under the Curve-Receiver Operating Characteristics Comparison By Classifiers

The AUC-ROC metric is a cornerstone in healthcare analytics, providing a clear and quantitative measure of how well a test or model performs in real-world clinical settings, thereby directly influencing patient diagnostics and treatment strategies.

ROC Curve (Receiver Operating Characteristic Curve): This curve plots the true positive rate (TPR, also known as sensitivity) against the false positive rate (FPR, 1 - specificity) at various threshold settings. The true positive rate is the proportion of actual positives correctly identified by the model, while the false positive rate is the proportion of actual negatives incorrectly classified as positives.

AUC (Area Under the Curve): The AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). This value provides a single number that describes the overall ability of the test to discriminate between those individuals who have the disease and those who do not. An AUC of 0.5 suggests no discrimination (equivalent to random guessing), and an AUC of 1.0 indicates perfect discrimination.

In our evaluation of the performance of various classifiers on AE-synthetic and original cardiovascular datasets using AUC scores, we observed that most models performed similarly on both datasets, with only slight variations in AUC scores. For instance, the Random Forest classifier achieved an AUC of 0.77 on AE-synthetic data and 0.78 on original data, suggesting consistent performance across different data types. The Gradient Boosting and Logistic Regression classifiers also displayed comparable performance with AUCs ranging from 0.77 to 0.81 across both datasets. This indicates a robust ability to generalize between synthetic and original data, which is crucial for models expected to perform in varied real-world scenarios. However, the MLP Classifier and XGB Classifier showed notable differences in performance between datasets, with a decrease in AUC when moving from original (AUCs

of 0.81) to synthetic data (AUCs of 0.73 and 0.74, respectively). This might suggest a sensitivity to the nuances of synthetic data generation techniques, which could impact the model's effectiveness in clinical settings where data variability is high. Conversely, the LGBM Classifier demonstrated superior performance on the original data with an AUC of 0.82 compared to 0.77 on the synthetic dataset, potentially indicating better handling of real-world data complexities. Lastly, both the KNN and Adaboost Classifiers showed a consistent decrement in performance on the synthetic dataset, suggesting that these models may require adjustments or re-tuning to better capture the characteristics of synthetic data.

Overall, these results underscore the importance of model selection based on the specific characteristics of the data used. The slight variations in AUC scores highlight the need for thorough testing on both synthetic and original datasets to ensure model robustness and reliability in different clinical environments. Figures depicting these performances can be found in Figure 4.64 and Figure 4.65.

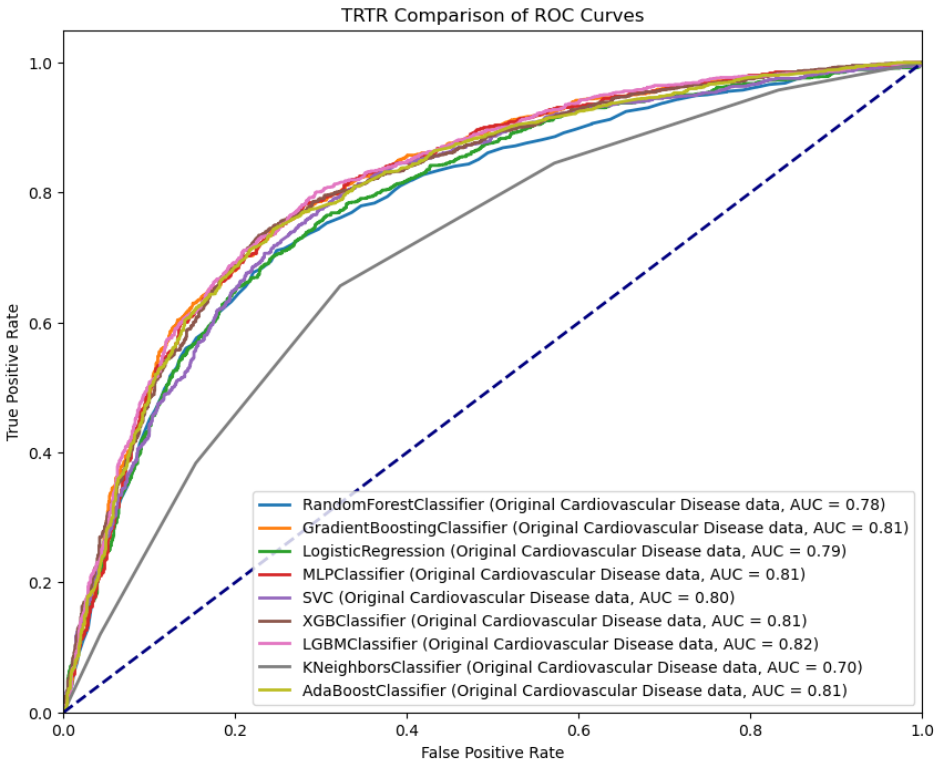


Figure 4.64: Comparison of ROC Curves for TRTR

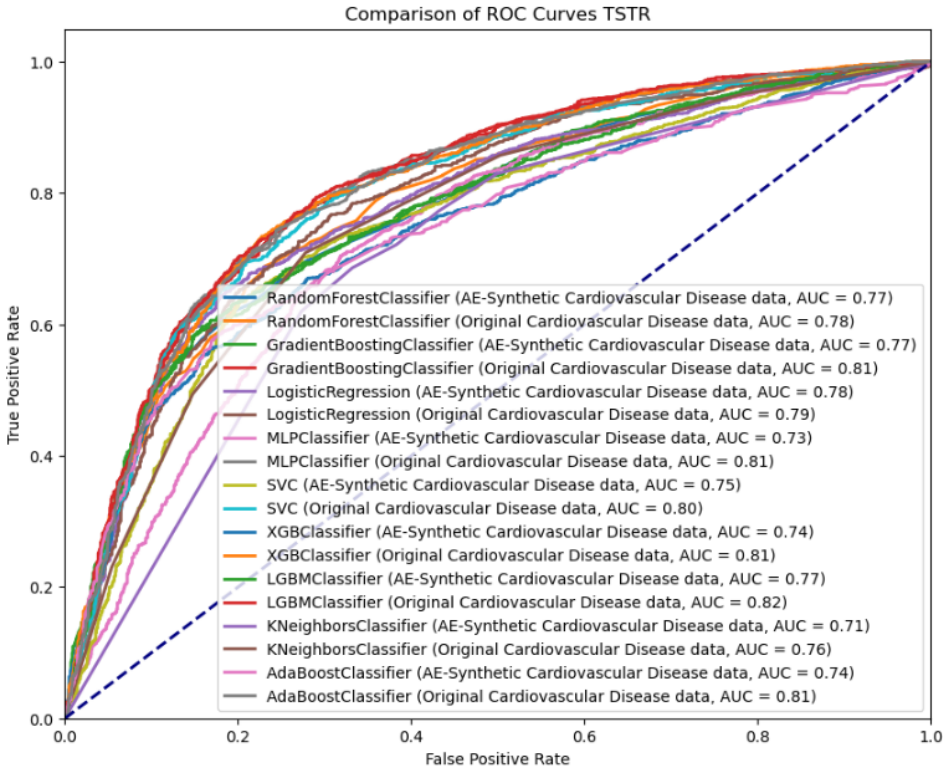


Figure 4.65: Comparison of ROC Curves Between TRTR and TSTR

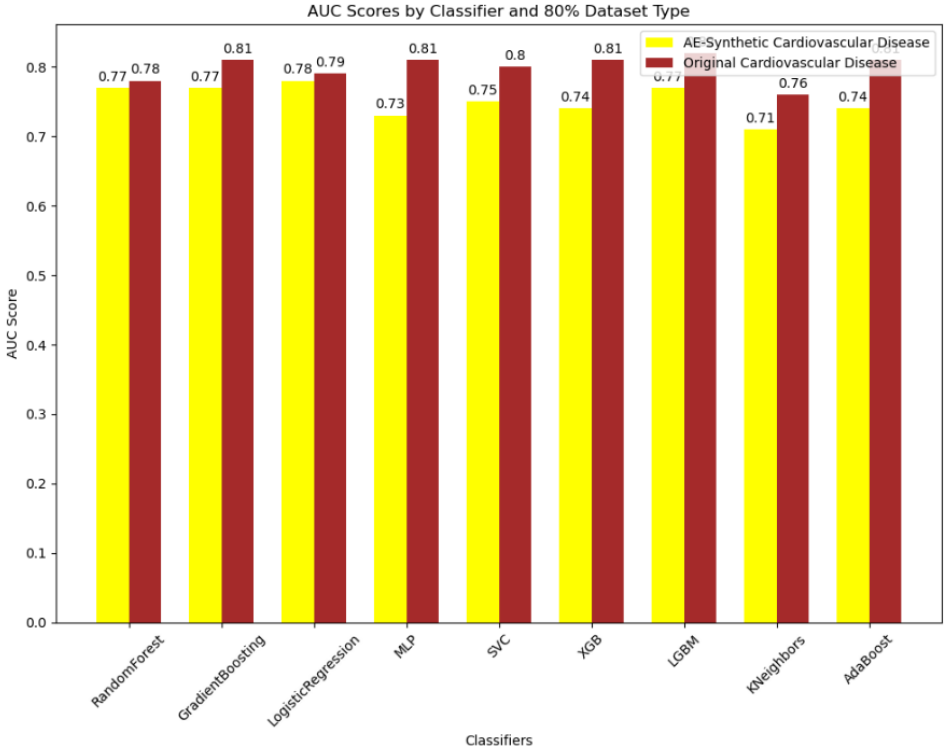


Figure 4.66: Area Under Curve Scores By Classifier Between 80% Original and AE-Synthetic Cardiovascular Disease Data - Back downward to Table 4.41

4.4.2 Comparative Analysis of Original and VAE Synthetic Cardiovascular Disease Data: A Multi-Faceted Evaluation Using VAE Model

In this section, we present the results of Chi-squared tests performed to examine the association between categorical features in the original and VAE-synthetic cardiovascular datasets. The Chi-squared test evaluates whether the distributions of categorical variables differ significantly between the two datasets.
As shown in Table 4.17, most features did not show significant differences (P-value > 0.05), suggesting that the VAE model has successfully captured the categorical distributions of the original data. Notable exceptions include 'Smoke' and 'Alcohol', where the P-values indicate significant differences, suggesting areas where the VAE model may not have perfectly replicated the original data's characteristics.
These findings are critical for validating the synthetic data's utility, particularly in ensuring that it can serve as a viable substitute for real data in sensitive applications without introducing biases or inaccuracies.

Table 4.17: Chi-Squared Test Results for Categorical Features

Feature	Chi-Squared Stat	P-value	Degrees of Freedom
Gender	0.433	0.510	1
Cholesterol	0.467	0.977	4
Glucose	4.890	0.299	4
Smoke	8.508	0.004	1
Alcohol	6.662	0.010	1
Active	0.000	1.000	0
Cardio	0.065	0.799	1

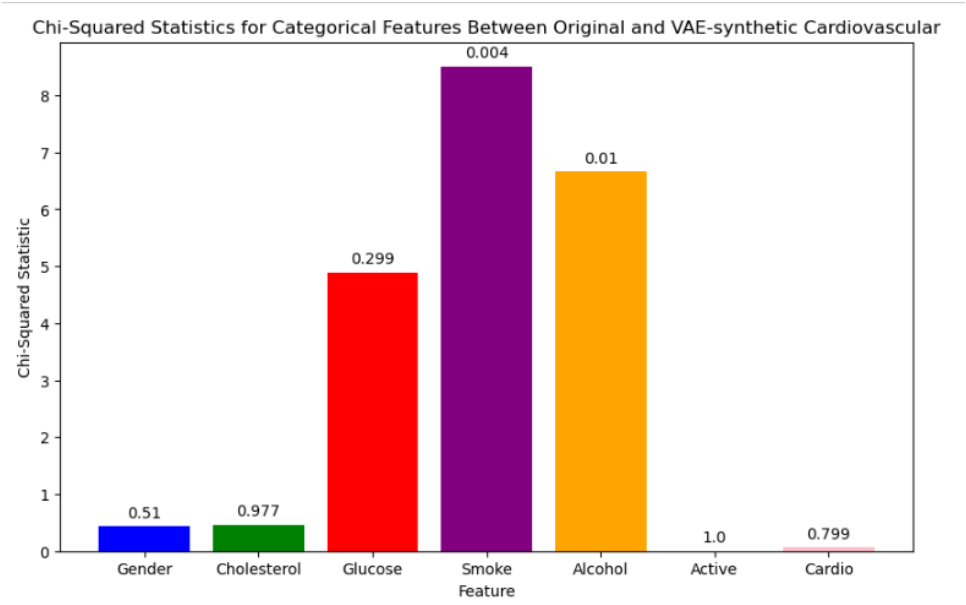


Figure 4.67: Chi-Squared Statistics for Categorical Features

Graphical representations further elucidate the results of our statistical tests. A bar chart visualizing the Chi-squared statistics, as depicted in Figure 4.67, clearly shows the comparative significance of each feature's test result. This visualization aids in quickly identifying features where the synthetic data varies significantly from the original, guiding further refinements to the synthetic data generation process.

Table 4.18: Error Metrics Comparing Original and VAE Synthetic Cardiovascular Data After Normalization

Feature	MSE	RMSE	MAE	Accuracy (%)
Age	0.0388	0.197	0.163	–
Gender	–	–	–	54.17
Height	0.0019	0.044	0.034	–
Weight	0.0068	0.082	0.064	–
Ap_hi	0.0001	0.011	0.0013	–
Ap_lo	0.0003	0.018	0.0029	–
Cholesterol	–	–	–	60.37
Glucose	–	–	–	76.99
Smoke	–	–	–	85.99
Alcohol	–	–	–	92.88
Active	–	–	–	80.38

Error Metrics Analysis

In our comprehensive evaluation of the performance differences between the original and VAE synthetic cardiovascular datasets, a series of error metrics were computed post-normalization to ensure comparability. Table 4.18 illustrates these metrics, showing significantly reduced errors in continuous features and high classification accuracy for categorical variables. This evidence supports the high fidelity of the synthetic data in replicating the statistical properties of the original dataset, thereby validating the effectiveness of our data synthesis approach. The reduced MSE, RMSE, and MAE values across key continuous variables such as age, height, and weight highlight the precision with which the synthetic dataset mimics the original data distribution. Notably, the accuracy metrics for categorical features like gender, cholesterol, and glucose confirm that the synthetic data maintains an accurate representation of binary and multinomial categories.

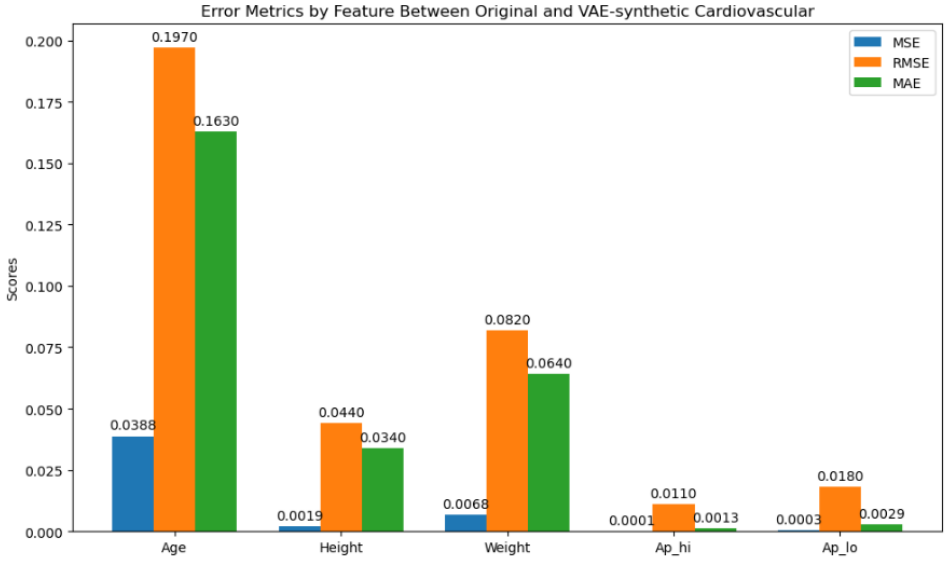


Figure 4.68: Comparative Analysis of Error Metrics and Accuracy between Original and VAE Synthetic Cardio Data

The bar graph 4.68 provides a visual comparison of three key error metrics across several features of a dataset, including Age, Height, Weight, Aphi (Systolic Blood Pressure), and Aplo (Diastolic Blood Pressure). The metrics displayed are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

MSE quantifies the average squared difference between the estimated values and the actual value, highlighting large errors due to its squaring of each term. RMSE, the square root of MSE, provides a measure of the magnitude of error in the same units as the original data, making it more interpretable. MAE measures the average magnitude of errors in a set of predictions, without considering their direction, presenting a linear score where all individual differences are weighted equally in the average. In the graph, the MSE values are shown in one color, emphasizing their respective magnitudes across the features. The RMSE values, displayed in another color, typically follow the trend in MSE but are scaled down as they are the square root of MSE values. The MAE values, also in a unique color, offer a direct measurement of average error magnitudes. Each bar's height indicates the metric's value for that particular feature, placed side by side for easy comparison. This visualization is particularly useful for identifying features with higher errors that may require further analysis or preprocessing to enhance model accuracy. The graph serves as a critical tool for diagnostic analysis in predictive modeling, helping to understand the distribution and severity of errors in model predictions across different features.

Correlation Matrix Comparison Between Original and VAE-Synthetic Cardiovascular Disease Data

The analysis of the numerical correlation matrices from the 80% Original and VAE-Synthetic Cardiovascular Disease data provides critical insights, shedding light on how feature relationships differ and resemble between the two datasets through visualization and direct comparison.

In the original dataset, age correlates positively with systolic and diastolic blood pressure (aphi and aplo) with values of 0.02 and 0.01 respectively, and negatively with height at -0.08, aligning with the common observations that blood pressure tends to rise and stature may decrease as people age. Gender shows strong correlations with height and smoking habits, with coefficients of 0.50 and 0.34, reflecting significant differences in physical attributes and lifestyle behaviors between genders. Additionally, cholesterol and glucose levels have a moderate correlation of 0.46, indicating a possible metabolic link relevant to cardiovascular health. Figure 4.69

In the VAE-Synthetic dataset, age correlates unusually high with almost all features, including notably strong correlations with aphi and aplo at 0.94 and 0.86, suggesting that the process of generating synthetic data might have overly accentuated the influence of age on other cardiovascular risk factors. Gender's correlation with height in the synthetic data is extremely high at 0.99, indicating a potential overfitting issue or an anomaly introduced during synthetic data creation. Similar to the original dataset, cholesterol and glucose maintain a strong correlation at 0.94, which is even more pronounced than in the original, possibly indicating that these metabolic interactions are either preserved or amplified in the synthetic version. Figure 4.70

The scatter plot in Figure 4.73 provides a visual comparison of the correlation coefficients for various features related to cardiovascular disease outcomes between the original and AE synthetic datasets. Each point on the plot corresponds to a feature, positioned according to its correlation coefficients in the original versus the synthetic dataset. A red dashed line illustrates perfect agreement, indicating that any point on this line shows an exact replication of the original dataset's correlation by the synthetic data. Points positioned above this line suggest a higher correlation in the synthetic dataset relative to the original, whereas points below the line suggest a lower correlation.

Key observations from the plot include a general close alignment between most features of the original and synthetic datasets, with most points clustering near or along the line of perfect agreement. Some features, such as 'height' and 'weight,' exhibit slight deviations, where the

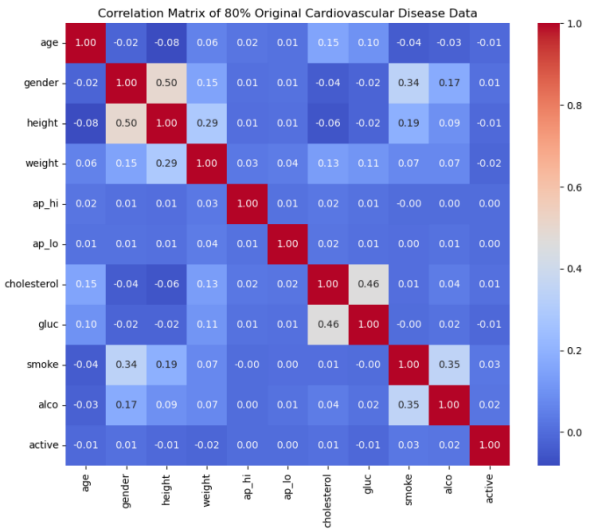


Figure 4.69: Original Cardiovascular Data

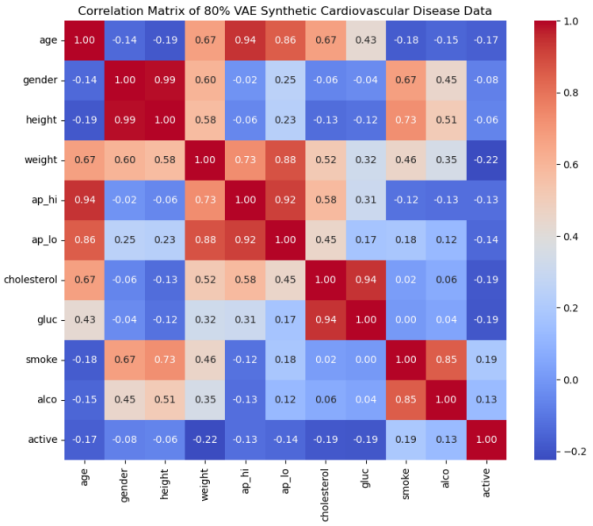


Figure 4.70: VAE Synthetic Cardiovascular Data

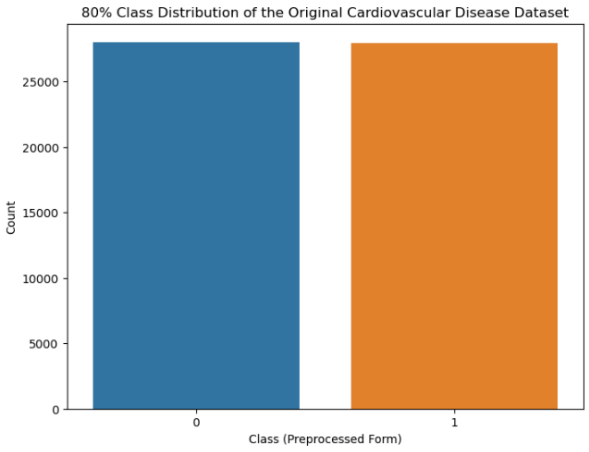


Figure 4.71: Original Cardio Disease Class Distribution

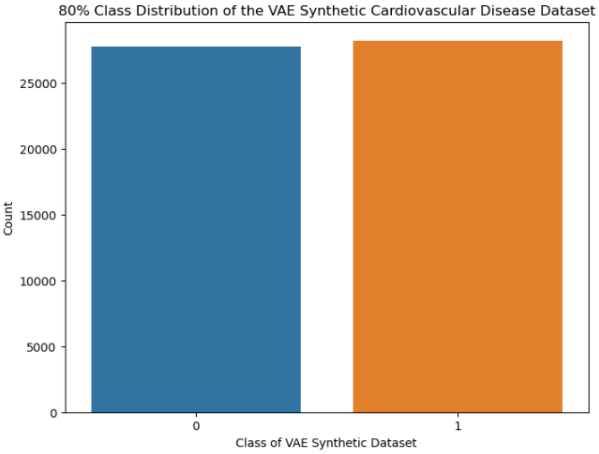


Figure 4.72: VAE-Synthetic Cardio Disease Class Distribution

synthetic data displays slightly altered correlation patterns compared to the original. Overall, the pattern observed suggests that the AE synthetic data retains a high degree of fidelity in mirroring the correlation structure of the original data. This fidelity is vital for the synthetic data's utility in applications such as modeling and risk prediction in healthcare. The graph is instrumental in evaluating the synthetic data's quality in maintaining statistical properties similar to the original, ensuring its appropriateness for use in sensitive sectors like healthcare.

The histogram in Figure 4.74 visually compares the correlation coefficients of various features with cardiovascular disease between the original and synthetic datasets. The bars represent how each feature correlates with the presence of cardiovascular disease in both datasets. From the graph, it's evident that most features maintain similar correlation magnitudes across both datasets, indicating that the synthetic data preserves the relational structure found in the original data. Notable features like **weight, height, cholesterol, smoke, and aplo** exhibit slight variations in correlation strength but generally align well between the two datasets.

This visual comparison helps to affirm the quality of the synthetic dataset in mimicking the original data's statistical properties, which is crucial for validating synthetic data applications in sensitive fields like healthcare. This ensures that the synthetic data can be used for

robust predictive modeling or statistical analyses without significant loss in the fidelity of the relationships within the data.

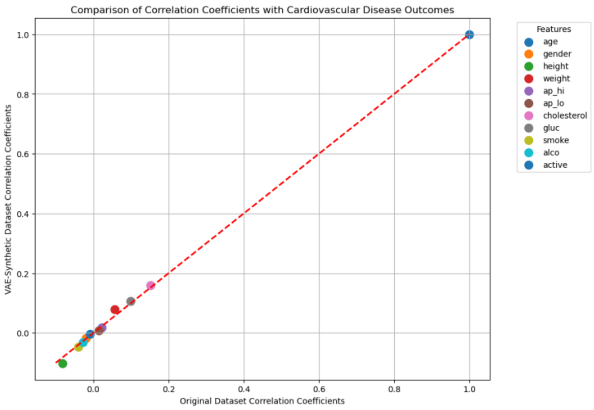


Figure 4.73: Comparison of Correlation Coefficients with Cardiovascular Disease Outcomes

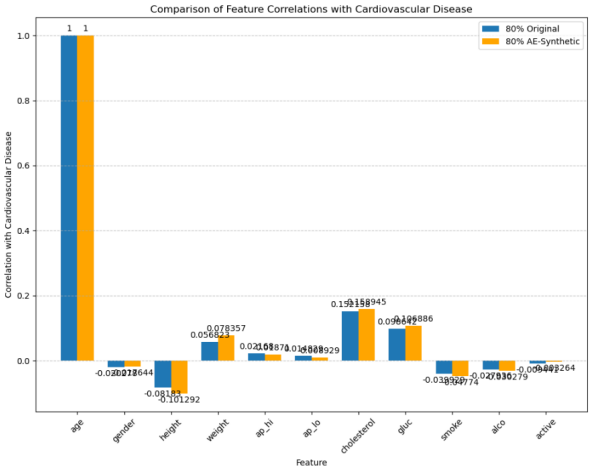


Figure 4.74: Comparison of Feature Correlations with Cardiovascular Disease

Area Under Curve Scores By Classifier Between Original and VAE-Synthetic Cardiovascular Disease Data

The ROC curve analysis for various classifiers on both VAE-synthetic and original cardiovascular disease data reveals distinct performance characteristics in Figure 4.75 and Figure 4.76. The synthetic data, generated through a VAE model, consistently achieves an Area Under the Curve (AUC) of 1.00 across all classifiers, indicating perfect discrimination ability between the classes. This perfect score suggests that the synthetic data preserves and perhaps even accentuates the defining characteristics of the dataset necessary for flawless classification. In contrast, classifiers trained on the original data show lower AUC values, with the highest being 0.82 for the LGBM classifier and the lowest at 0.70 for the KNN classifier. These values, while respectable, highlight the inherent noise and complexity of real-world data that synthetic data might not fully replicate. The Gradient Boosting and SVC classifiers show relatively strong performance on the original data with AUC scores of 0.81 and 0.80, respectively, indicating robustness in handling the original data's variability. Figure 4.77

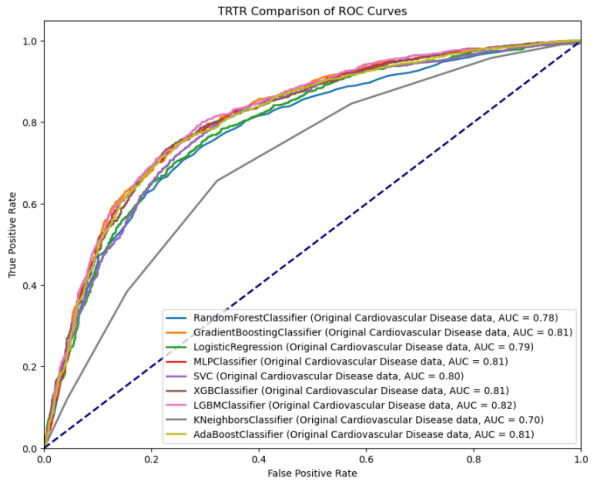


Figure 4.75: AUC-ROC Curve for TRTR

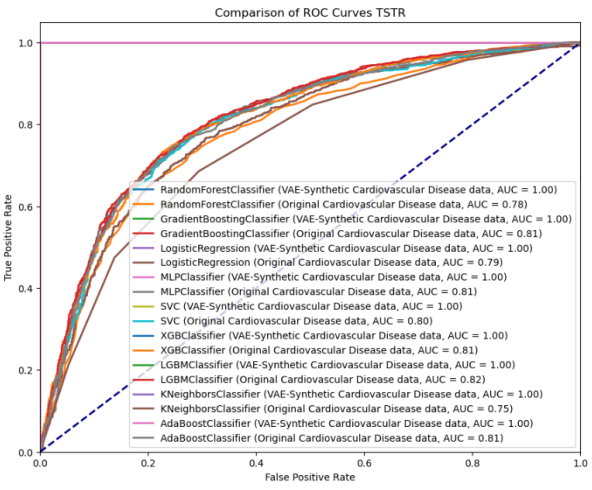


Figure 4.76: Comparison of ROC Curves Original and VAE-Synthetic (TRTR and TSTR)

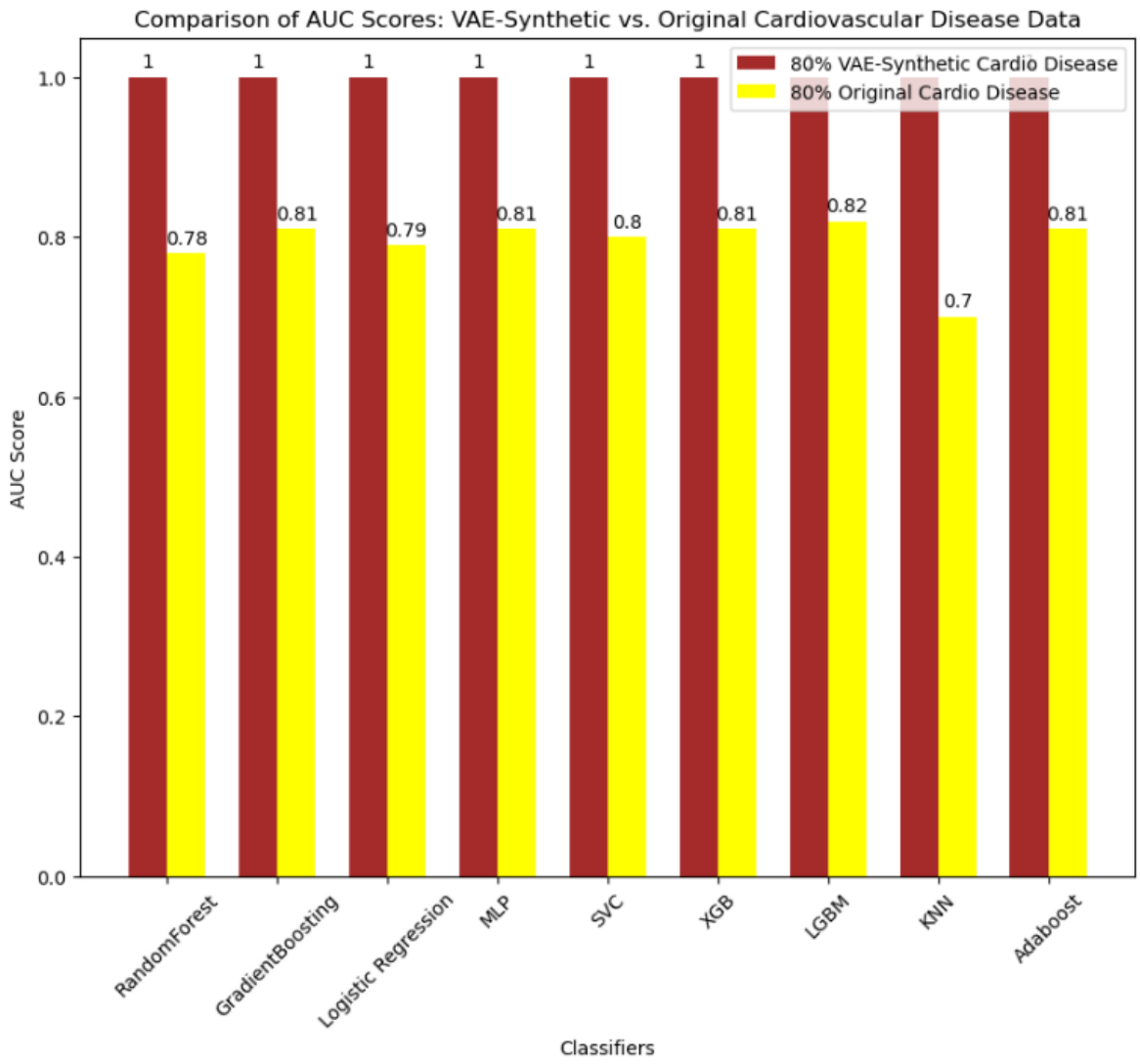


Figure 4.77: Area Under Curve Scores By Classifier Between 80% Original and VAE-Synthetic Cardiovascular Disease Data - Back downward to Table 4.41

Comparison of Cross-Validation Accuracy Between Original and Variant VAE-Synthetic Cardiovascular Disease Data

The bar graph in Figure 4.78 provides a detailed comparison between the performance of classifiers on the original cardiovascular disease data and the VAE synthetic data across different configurations of training epochs and dense layer counts. This graphical representation is key to assessing the robustness and reliability of the VAE synthetic data, as it vividly illustrates the consistently high cross-validation accuracy scores near perfection (1.0000) achieved by the synthetic data. These results are juxtaposed against the significantly lower accuracy scores from the original data, which vary more broadly across classifiers.

The graph includes five sets of bars for each classifier, showing classifier performance at 15000 epochs with 64 layers, 5000 epochs with 32 layers, 1000 epochs with 32 layers, and 500 epochs with 16 layers for the synthetic data, alongside their original data counterparts in grey. The consistently high scores in the synthetic setups underscore the effectiveness of the VAE model in creating highly accurate, yet potentially overfitting, synthetic replicas of the original data. In stark contrast, the original data's performances are markedly lower, highlighting potential overfitting issues with the synthetic models given their near-perfect accuracy.

These visual findings are crucial for understanding not only the quality of the synthetic data

but also the potential privacy implications. The near-perfect performance across different synthetic configurations suggests a strong replication of the original data's structure and patterns, possibly at the expense of privacy, as indicated by the high accuracy potentially capturing too much detail from the original data.

Tables in Table 4.19 complement these graphs by providing exact numeric values of accuracies and standard deviations, allowing for precise comparisons and a deeper understanding of model behavior across different data sets and configurations. This comprehensive presentation aids stakeholders in making informed decisions about the utility and risks associated with using synthetic data in various applications, especially in sensitive fields like healthcare.

Table 4.19: Cross-Validation Accuracy of VAE-Synthetic Cardiovascular Disease Data Across Different Configurations

Classifier	**15000 Epoc, 64-Dens**	**5000 Epoc, 32-Dens**	**1000 Epoc, 32-Dense**	**500 Epoc, 16-Dense**
DCT	1.00 (±0.00)	1.00 (±0.00)	0.99 (±0.01)	0.99 (±0.03)
GDB	1.00 (±0.00)	1.00 (±0.00)	0.99 (±0.02)	0.99 (±0.03)
RDF	1.00 (±0.00)	1.00 (±0.00)	1.00 (±0.00)	0.99 (±0.02)
AdaBoost	1.00 (±0.00)	1.00 (±0.00)	0.99 (±0.01)	0.99 (±0.01)
LGBM	1.00 (±0.00)	1.00 (±0.00)	0.99 (±0.01)	0.99 (±0.01)
XGB	1.00 (±0.01)	1.00 (±0.01)	1.00 (±0.01)	0.99 (±0.02)
KNN	1.00 (±0.00)	1.00 (±0.00)	1.000 (±0.00)	0.99 (±0.01)
LGR	1.00 (±0.00)	1.00 (±0.00)	0.99 (±0.01)	0.99 (±0.01)
SVC	1.00 (±0.00)	1.00 (±0.00)	1.000 (±0.00)	0.99 (±0.01)
MLP	1.00 (±0.00)	1.00 (±0.00)	1.00 (±0.01)	1.00 (±0.00)

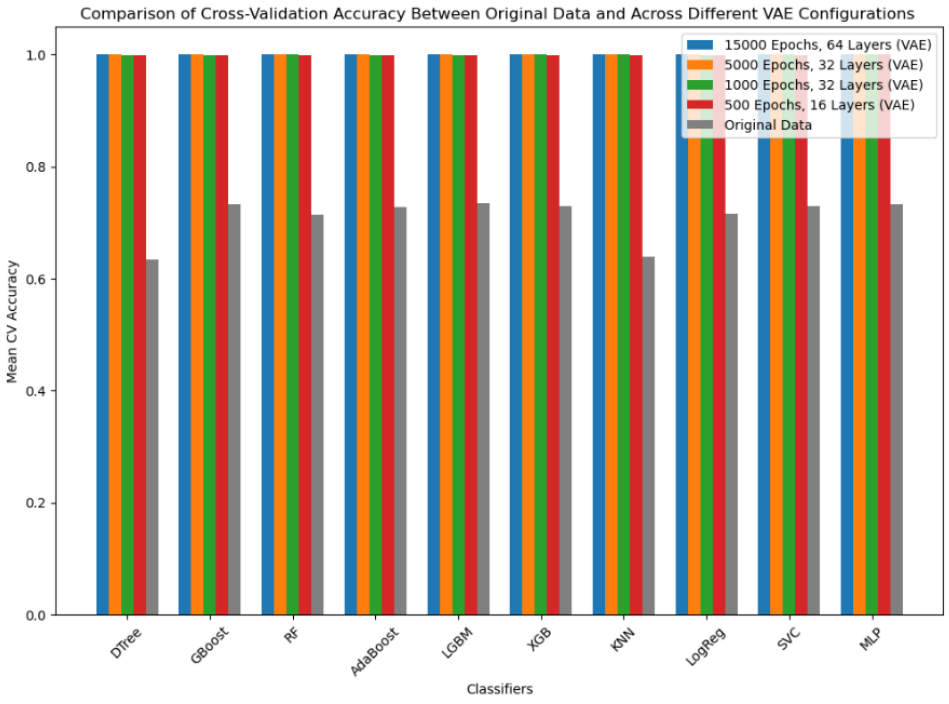


Figure 4.78: Comparison of Cross-Validation Accuracy Between 80% Original and Across Variant VAE Configurations - Back downward to Table 4.41

4.4.3 Comparative Analysis Between AE-Synthetic and VAE-Synthetic Cardiovascular Disease Datasets

Statistical Independence and Chi-Squared Tests

The statistical independence of categorical features in both AE and VAE synthetic datasets was assessed using Chi-squared tests, comparing each dataset against the original data.
AE Dataset: As detailed in Table 4.16, the AE synthetic data showed significant discrepancies, particularly for Glucose and Cholesterol, with P-values of 0.0331 and 0.0329 respectively, indicating a substantial divergence from the expected distributions.
VAE Dataset: The VAE synthetic data, in contrast, showed broader alignment across most features as shown in Table 4.17, except for Smoke and Alcohol where P-values were significantly lower, indicating divergence in these areas.

Error Metrics and Data Fidelity

Error metrics such as MSE, RMSE, and MAE were calculated to evaluate the precision of the synthetic datasets in mimicking the original dataset's continuous variables.
AE Dataset: Table 4.16 presents the AE dataset's error metrics, which indicate slightly higher values, suggesting less precision in replicating the original data's continuous features.
VAE Dataset: The VAE dataset's error metrics, shown in Table 4.18, demonstrate significantly reduced errors, indicating a higher fidelity to the original data. This table also highlights high classification accuracy for categorical features, emphasizing the VAE's effective data characteristic preservation.

Correlation Matrix Analysis

Analysis of the correlation matrices from both synthetic datasets provides insights into how well each dataset preserved the relationships between features compared to the original data.
AE Synthetic Data: The correlation matrix for the AE data (Figure 4.57) sometimes showed exaggerated correlations, suggesting potential overfitting.
VAE Synthetic Data: The correlation matrix for the VAE data (Figure 4.70) displayed more balanced correlations that closely mirrored the original data, indicating a more accurate and unbiased representation.

Classifier Performance on Synthetic Data

The performance of classifiers on these datasets offers insights into each dataset's utility for training machine learning models.
AE Synthetic Data: Classifiers generally showed reduced performance on the AE synthetic dataset, suggesting issues with the dataset's generalizability.
VAE Synthetic Data: The VAE dataset (refer to the graphical representations in Figure 4.68) provided closer performance metrics to the original data across classifiers, highlighting its effectiveness for predictive analytics.

Visual and Graphical Comparisons

Graphical representations provide a visual comparison of statistical tests and error metrics between the datasets.
AE Visuals: Figure 4.54 illustrates the AE dataset's Chi-Squared statistics, showing greater deviations from the original data.
VAE Visuals: Figure 4.67 provides a clearer picture of the VAE dataset's closer alignment with the original data, both in statistical tests and error metrics.

This detailed comparative analysis highlights the VAE synthetic dataset's superiority in

replicating the original cardiovascular disease data across various dimensions. The VAE dataset's closer alignment with the original data in key aspects such as statistical tests, error metrics, and classifier performance makes it a more suitable choice for robust applications in healthcare analytics. This comparison, underpinned by specific references to tables and figures, enhances the clarity and utility of the findings, aiding stakeholders in making informed decisions about synthetic data usage in sensitive healthcare applications.

4.5 Privacy of Cardiovascular Disease Data

4.5.1 Evaluation of Privacy Preservation through Singling-Out Univariate Risk Assessment on AE Synthetic Cardiovascular Data

Table 4.20: Tabular Representation of Singling-Out Univariate Risk Assessment on 80% AE Synthetic Cardiovascular Data

Evaluation Metric	**n_attacks=1500**	**n_attacks=500**
Main Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Baseline Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Control Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Privacy Risk	0.0 (CI: 0.0, 0.0018); 0.0 (CI: 0.0, 0.0054)	

Table 4.20 includes the results from both attack scenarios (1500 and 500 attacks), detailing the Privacy Risk values, Confidence Intervals, and success rates for main, baseline, and control attacks. The privacy risk evaluation for the AE synthetic cardiovascular data was conducted using the Singling Out Evaluator, assessing the likelihood that an individual could be uniquely identified within the dataset. This evaluation was performed under two different attack scenarios—1500 and 500 total attacks—focusing on univariate analysis.

- **Singling Out Risk via Univariate Analysis (1500 Attacks):** The assessed privacy risk was remarkably low, with a privacy risk value of 0.0 and a confidence interval extending up to approximately 0.0018. This indicates an extremely low probability of singling out any individual from the ae synthetic dataset based on univariate attributes. This is seen in Table 4.20.

- **Main, Baseline, and Control Attack Success Rates (1500 Attacks):** All three metrics showed identical success rates at approximately 0.0013. The error associated with these rates was also the same, underscoring the consistency and reliability of the attack model used in this scenario. Can be seen in Table 4.20.

- **Singling Out Risk via Univariate Analysis (500 Attacks):** Similar to the first scenario, the privacy risk remained at 0.0, but with a slightly higher confidence interval upper bound at approximately 0.0054. This increase is minimal but suggests a slight rise in potential risk when fewer attacks are conducted. This is found in Table 4.20.

- **Main, Baseline, and Control Attack Success Rates (500 Attacks):** In this scenario with fewer attacks, the success rates were slightly higher, at approximately 0.0038. The error remained consistent with the value of the success rate, indicating that the increase in risk is detectable but still minimal. Found in Table 4.20.

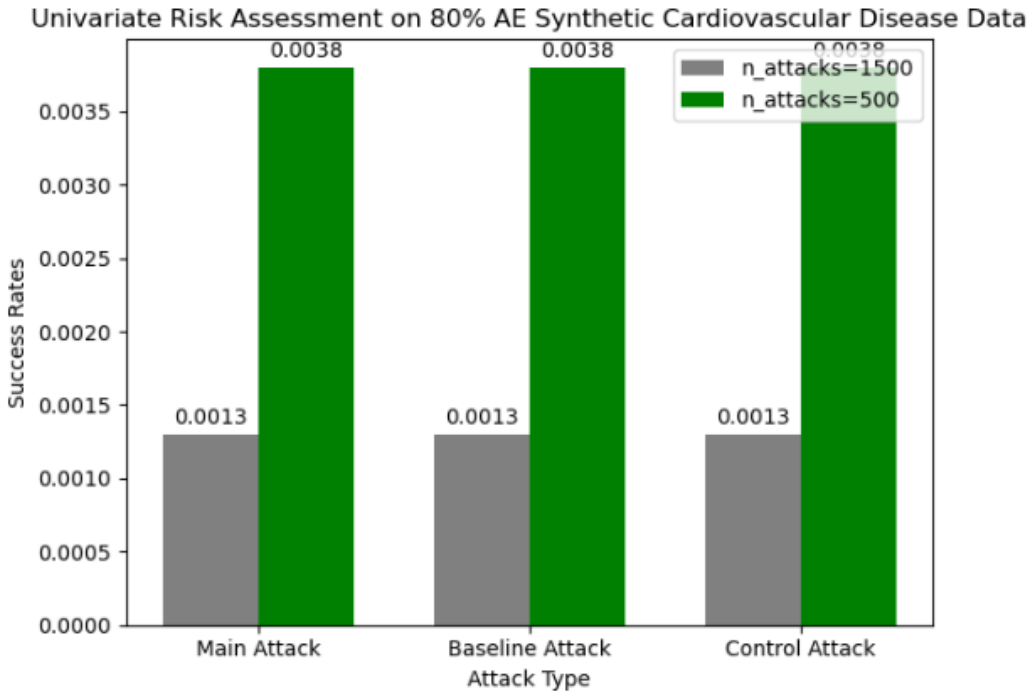


Figure 4.79: Univariate Risk Assessments Success Rates by Number of Attacks on 80% AE Synthetic Cardiovascular Disease Data

The bar chart in Figure 4.79 illustrates the success rates of different types of attacks conducted with two different numbers of attempts (1500 and 500 attacks). The success rates are notably low across all attack types, with all three categories—Main, Baseline, and Control attacks—showing identical success rates of 0.0013 for 1500 attacks and 0.0038 for 500 attacks. This uniformity suggests that the number of attacks significantly influences the likelihood of success, though the overall effectiveness remains minimal. The increased success rate with fewer attacks could imply a nuanced vulnerability that appears more often under less frequent testing conditions.

The pie charts in Figure 4.80 illustrate the success versus failure rates for privacy risk assessments conducted with 1500 and 500 attack attempts, respectively, as well as an overall comparison:

This chart in Figure 4.80 demonstrates a very low success rate of 0.13% (Grey) and a failure rate of 99.87% (Green), indicating strong privacy preservation against singling-out attacks when tested with 1500 attempts. Similarly, this chart shows a slightly higher success rate for 500 attacks at 0.38(or 0.4%) (Coral) than the 1500 attempts and a failure rate of 99.62% (Light Blue). Still, it remains significantly low, reinforcing the robustness of the privacy measures in place. The pie chart aggregates the success rates across all attacks to provide an overarching view of the effectiveness of these privacy attacks against the dataset. With a combined success rate approximating 0.26 (0.3%) (Gold) the vast majority of attempts to compromise data privacy result in failure approximately 99.7% of the time (Grey), underscoring the robustness of the data protection mechanisms in place. This visual emphasizes the effectiveness of current privacy-preserving methods implemented in safeguarding synthetic data against potential unauthorized singling out of individuals.Figure 4.80 These visualizations provide a clear and concise presentation of the privacy risk assessment, underlining the effectiveness of the synthetic data in preventing successful singling-out attacks and maintaining the confidentiality of the underlying original data.

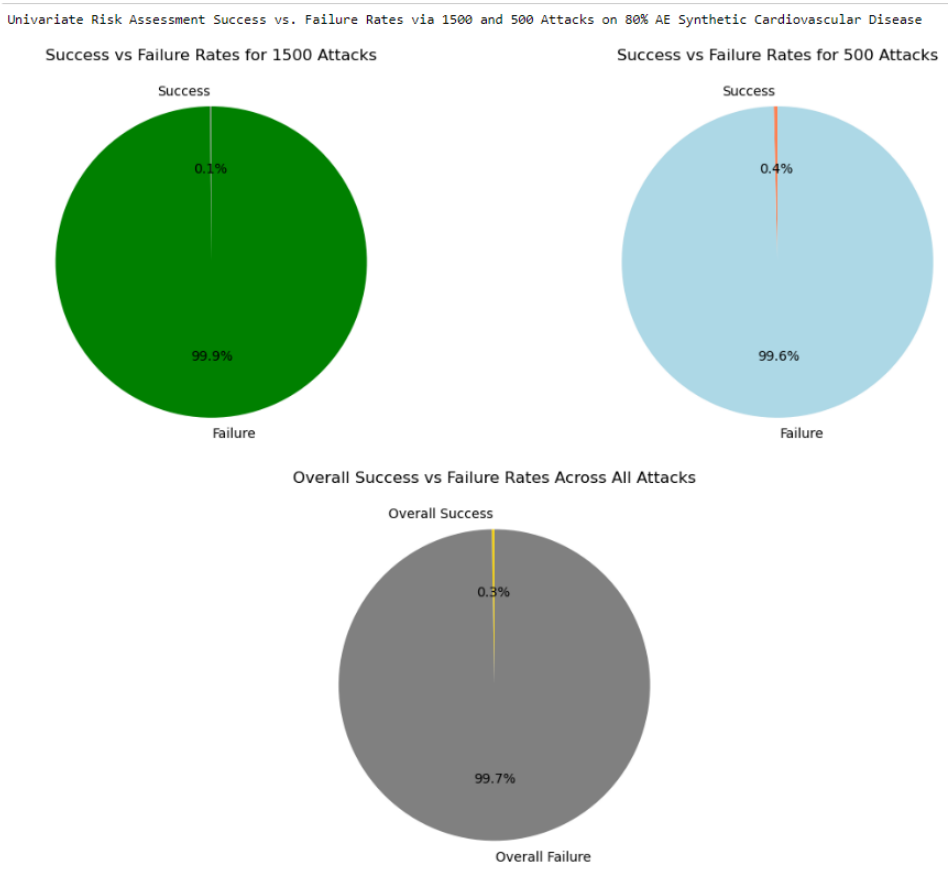


Figure 4.80: Univariate Risk Assessments Success/Overall Success vs. Failure Rates by Number of Attacks on 80% AE Synthetic Cardiovascular Disease Data

4.5.2 Evaluation of Privacy Preservation through Singling-Out Multivariate Risk Assessment on AE Synthetic Cardiovascular Data

Table 4.21: Multivariate Singling-Out Risk Assessment on AE Synthetic Cardiovascular Data

Evaluation Metric	**n_attacks=1500**	**n_attacks=500**
Main Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Baseline Attack Success Rate	0.0019 (±0.0018)	0.0038 (±0.0038)
Control Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Privacy Risk	0.0 (CI: 0.0, 0.0018); 0.0 (CI: 0.0, 0.0054)	

In the conducted privacy risk multivariate assessments through singling-out evaluations of the autoencoder-generated synthetic cardiovascular dataset, several key insights emerged:

- **Multivariate Analysis (1500 Attacks):** The privacy Risk is almost negligible with a value of 0.0, indicating no significant risk within a 95% confidence interval (CI) ranging up to 0.0018. In the **Main Attack**, the success rate of this attack was minimal at 0.0013 (±0.0013), reflecting a very low probability of successfully singling out an individual. The **Baseline and Control Attacks** both showed identical success rates to the main attack, reinforcing the consistency and robustness of the dataset against such privacy threats. This can be fact-checked in Table 4.21
- **Multivariate Analysis (500 Attacks):** Similarly, a privacy risk of 0.0 was reported for 500 attacks, with a slightly wider CI of up to 0.0054 due to the reduced number of attacks, indicating a still robust defense against privacy breaches. The success rates for all types of attacks, including the main, baseline, and control, were identically 0.0038

(±0.0038). Despite the increased rate compared to the 1500 attack scenario, the risk remains markedly low. Table 4.21.

This bar chart in Figure 4.81 distinctly showcases the relative success rates of different attacks under varying numbers of attempts, indicating a consistent but minor increase in success rate with fewer attacks, suggesting that larger attack volumes slightly dilute the attack's effectiveness due to the robust synthetic data generation techniques used.

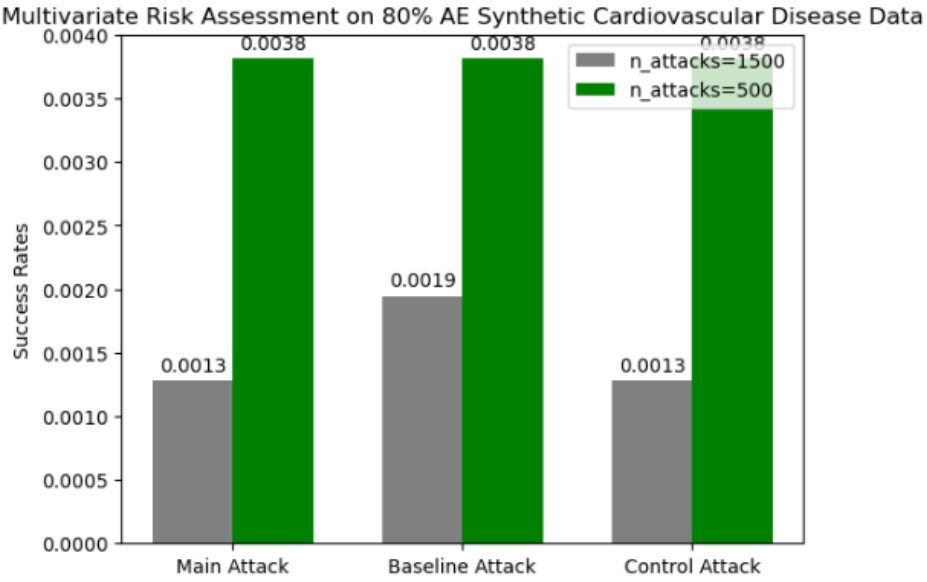


Figure 4.81: Multivariate Risk Assessments Success Rates by Number of Attacks on 80% AE Synthetic Cardiovascular Disease Data

The pie charts in Figure 4.82 display the success versus failure rates for multivariate singling-out attacks conducted with different numbers of attempts on ae-synthetic cardiovascular data. It conveys the low success rates of attacks, underscoring the synthetic dataset's efficacy in maintaining the privacy of individual data points. Here are the details:

- **1500 Attacks Success vs. Failure:** This chart shows a minimal success rate at approximately 0.1(0.13%) (grey) and a failure rate at 99.87%(green) for 1500 attacks, indicating that the AE synthetic data provides strong privacy protection against such attacks. Figure 4.82

- **500 Attacks Success vs. Failure:** The success rate for 500 attacks is slightly higher at around 0.4(0.38%) (gold) and a failure rate at 99.62% (light blue), but still remains significantly low, reinforcing the robustness of the ae-synthetic data in maintaining privacy. Figure 4.82

- **Overall Attack Success vs. Failure:** The overall pie chart aggregates the success rates from both attack scenarios, showing a combined success rate that remains below 0.4% (yellow), suggesting effective privacy preservation across multiple attack intensities. Figure 4.82

  This composite view emphasizes the robustness of privacy-preserving measures in the AE synthetic cardiovascular dataset, reinforcing the low likelihood of successful attacks across scenarios. This chart gives a general sense of the dataset's resistance to singling-out attacks, showcasing its potential for secure and privacy-preserving applications.

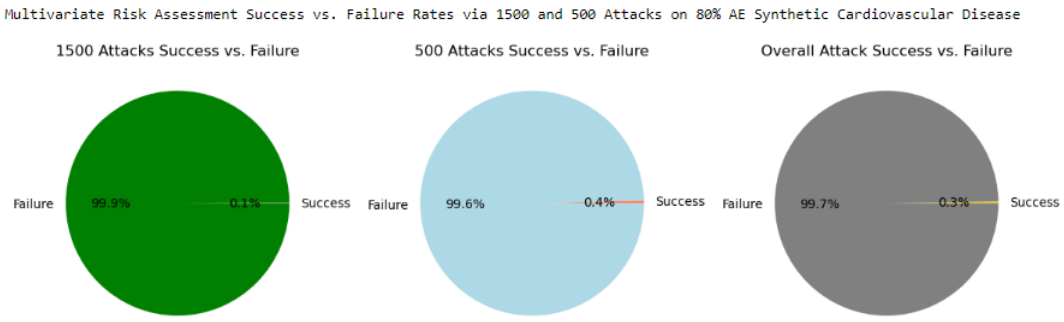


Figure 4.82: Multivariate Risk Assessments Success vs. Failure Rates by Number of Attacks on 80% AE Synthetic Cardiovascular Disease Data

4.5.3 Evaluation of Privacy Preservation through Linkability Risk Assessment on AE Synthetic Cardiovascular Data

In the assessment of linkability risks within the AE synthetic cardiovascular dataset, a detailed evaluation was conducted using a linkability evaluator, which measured the potential for matching individual records between the synthetic dataset and a control set. The evaluations were carried out under two conditions, with varying neighbor counts (nneighbors=10 and nneighbors=5) to test the robustness of linkability under different complexity levels.

- **Linkability Evaluation with 10 Neighbors:** This linkability risk was quantified at a minimal value of 0.000358, with a confidence interval stretching from 0.0 to 0.001357. This indicates a very low probability of correctly linking data records based solely on the available attributes. The success rates of the attacks were also low, with the main attack achieving a 0.002065 success rate, further demonstrating the difficulty of linking records accurately in this setting.
- **Linkability Evaluation with 5 Neighbors:** This linkability risk was reported as zero, reinforcing the strength of data anonymization in the synthetic dataset. The confidence interval remained negligible at 0.000432 at the upper bound. Success rates of attacks in this setting were slightly lower than in the 10 neighbor setting, with the main attack posting a success rate of 0.000494, showcasing an even stronger resistance to linkability.

Table 4.22: Linkability Risk Assessment for AE Synthetic Cardiovascular Data

Evaluation Metric	**n_neighbors=10**	**n_neighbors=5**
Main Attack Success Rate	0.0021 (±0.0007)	0.0005 (±0.0003)
Baseline Attack Success Rate	0.0016 (±0.0007)	0.0006 (±0.0004)
Control Attack Success Rate	0.0017 (±0.0007)	0.0006 (±0.0004)
Privacy Risk	0.0004 (CI: 0.0, 0.0014)	0.0 (CI: 0.0, 0.0004)

Table 4.22, details the outcomes of linkability risk assessments performed with two distinct configurations of neighbor counts (10 and 5 neighbors). The metrics presented in the table are:

- **Main Attack Success Rate:** Shows the proportion of successful main attacks conducted against the dataset, indicating the ease with which individual records can be linked back to their sources.
- **Baseline Attack Success Rate:** Illustrates the success rate of baseline attacks, providing a standard to measure the effectiveness of the main attack.
- **Control Attack Success Rate:** Details the success rate of control attacks, which typically involve a different or random subset of data, used to contextualize the main attack results.

- **Privacy Risk:** Represents the calculated risk of an individual's information being correctly re-identified, with a confidence interval (CI) provided for transparency and reliability.

This table effectively encapsulates the quantitative measures of privacy risks associated with synthetic data, illustrating the data's resilience or vulnerability to different types of linkability attacks. It serves as a critical resource for researchers and practitioners to evaluate the effectiveness of data anonymization techniques and the inherent privacy risks.

The bar graph in Figure 4.83 illustrates the success rates of three types of linkability attacks—main, baseline, and control—evaluated under two different configurations of neighbor counts (nneighbors=10 and nneighbors=5). Here's a detailed breakdown: The **Main Attack** which shows a slight decrease in success rate when the number of neighbors is reduced from 10 to 5, suggesting that the proximity of data points can impact the effectiveness of linkability. The **Baseline Attack** follows a similar pattern to the main attack, with a minor change in success rate, further emphasizing the robustness of the synthetic dataset against standard attack models. And, the **Control Attack** exhibits an almost constant success rate, indicating a consistent resistance across different neighbor configurations. This bar graph serves as a quantitative assessment tool that highlights the impact of neighborhood size on the potential for linkable data extraction, showcasing the synthetic dataset's capability to maintain low linkability risks across varying analytical conditions.

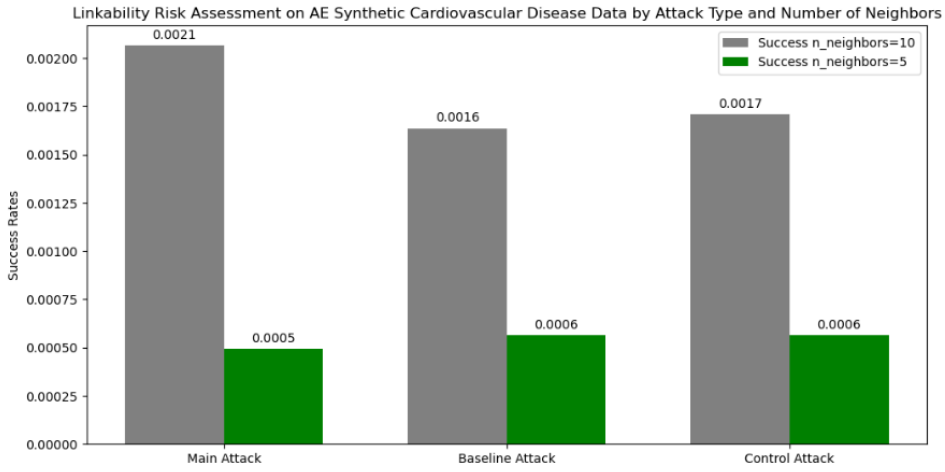


Figure 4.83: Linkability Risk Assessments Success Rates by Number of Neighbors on 80% AE Synthetic Cardiovascular Disease Data

The pie charts in Figure 4.84 represents the success versus failure rates of linkability attacks for two settings based on the number of neighbors (nneighbors = 10 and nneighbors = 5), along with an overall combined view: where success is defined by the ability to link data points in the synthetic dataset back to the original data.

- **10 Neighbors Success vs. Failure Rates:** This chart illustrates the success rates of different types of attacks (main, baseline, and control) when considering the 10 nearest neighbors in the dataset. The success rates are relatively low (0.2%), indicating that the AE synthetic cardiovascular data maintains a high level of privacy, making it difficult to link specific data points to individuals when a larger neighborhood is considered. Figure 4.84
- **5 Neighbors Success vs. Failure Rates:** With a smaller neighborhood of 5 neighbors, the success rates slightly increase, further affirming the effectiveness of the AE synthetic data in protecting against linkability attacks. This setting reveals an even better protection mechanism due to the closer proximity of the considered neighbors, yet the success rates remain significantly low. Figure 4.84

- **Overall Success vs. Failure Rates:** The combined pie chart aggregates the success rates from both neighbor settings, providing a holistic view of the effectiveness across different configurations. This overall perspective emphasizes the robustness of the synthetic data against linkability attacks across varied settings. Figure 4.84

  Each chart displays a predominant portion representing the 'Failure' rate, highlighting the substantial protection offered by the AE synthetic dataset against linkability attacks. These visual representations are crucial for understanding the effectiveness of privacy-preserving measures in synthetic datasets, indicating strong resistance against potential privacy breaches.

Figure 4.84: Linkability Risk Assessments Success vs. Failure Rates by Number of Neighbors on 80% AE Synthetic Cardiovascular Disease Data

4.5.4 Evaluation of Privacy Preservation through Inference Risk Assessment Per-Column on AE Synthetic Cardiovascular Data

The Inference Risk Assessment for AE Synthetic Cardiovascular Data highlights critical insights into the susceptibility of various health attributes to inference attacks, utilizing the smallest dataset size available as the basis for the number of attack attempts. Notably, the attribute **height** displayed a notable risk with a measured inference risk value of 0.0296 and a confidence interval ranging from 0.0043 to 0.0549, suggesting a moderate vulnerability. Similarly, **weight and aphi** (arterial pressure, high) showed measurable but lower risks, with values of 0.0135 and 0.0085 respectively. Other attributes like **age, gender, aplo, cholesterol, gluc, smoke, alco, and cardio** demonstrated minimal to no inference risk, indicating a strong protection level or lesser relevance in inference attack susceptibility.

Table 4.23 provides a concise overview of the vulnerability of each attribute in the dataset, helping researchers understand where the synthetic dataset stands in terms of privacy protection. **Attribute** lists the different data columns (attributes) assessed for inference risk. **Privacy Risk** shows the measured risk value for each attribute, indicating how likely it is that the attribute can be inferred. **Confidence Interval** provides the range within which the true inference risk is expected to fall, offering statistical reliability. And, **Success Rate (Main Attack)** indicates the effectiveness of the primary attack method in inferring the attribute, with the uncertainty expressed as a margin of error.

The bar graph in Figure 4.85 representation aids in visually delineating which attributes are more prone to privacy breaches, with **height** showing a significantly higher risk than most other attributes, followed by **weight and aphi.** The success rates of the attacks further substantiate the assessment: **Main Attack** achieved a success rate of 53.19%, signifying a high likelihood of attack success across most attributes. **Baseline Attack** Recorded slightly lower success at 49.95%, indicating the baseline scenario's comparative difficulty. **Control Attack**

Table 4.23: Inference Risk Assessment on AE Synthetic Cardiovascular Data

Attribute	Privacy Risk	Confidence Interval	Success Rate (Main Attack)
Age	0.0	(0.0, 0.0063)	53.2% (±0.83%)
Gender	0.0	(0.0, 0.0050)	49.9% (±0.83%)
Height	0.0296	(0.0043, 0.0549)	54.7% (±0.82%)
Weight	0.0135	(0.0, 0.0400)	53.2% (±0.83%)
Ap_hi	0.0085	(0.0, 0.0304)	49.9% (±0.83%)
Ap_lo	0.0	(0.0, 0.0)	54.7% (±0.82%)
Cholesterol	0.0	(0.0, 0.0002)	53.2% (±0.83%)
Gluc	0.0019	(0.0, 0.0174)	49.9% (±0.83%)
Smoke	0.0006	(0.0, 0.0033)	54.7% (±0.82%)
Alco	0.0	(0.0, 0.0002)	53.2% (±0.83%)
Active	0.0	(0.0, 0.0218)	49.9% (±0.83%)
Cardio	0.0	(0.0, 0.0)	54.7% (±0.82%)

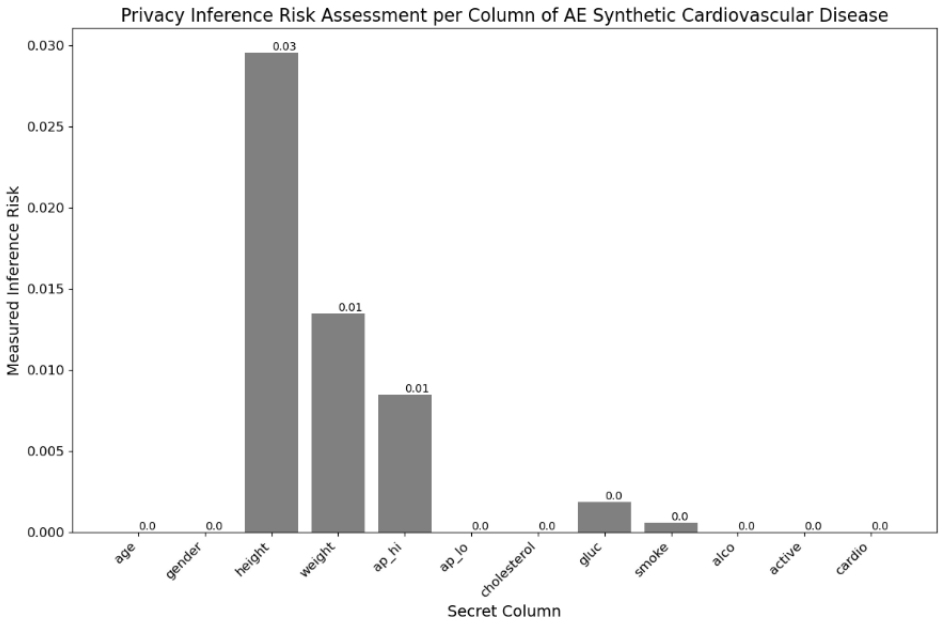


Figure 4.85: Inference Risk Assessments Success Rates by smallest size dataset on 80% AE Synthetic Cardiovascular Disease Data

Mirrored the main attack closely with a 54.70% success rate, suggesting similar vulnerability levels in both the synthetic and control setups.

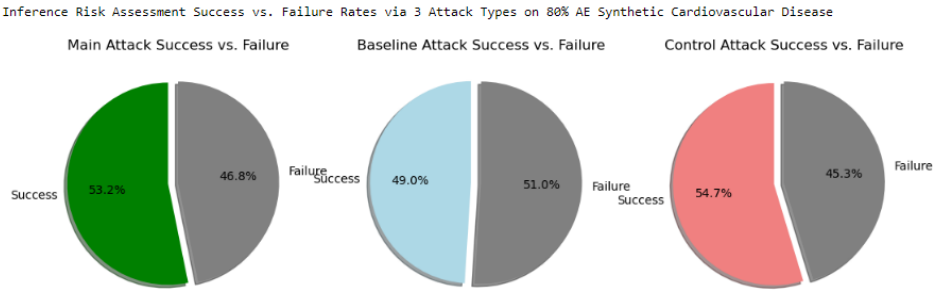


Figure 4.86: Inference Risk Assessments Success vs. Failure Rates by Size of smaller dataset on 80% AE Synthetic Cardiovascular Disease Data

The pie charts in Figure 4.86 depicts the success versus failure rates for each type of attack: **Main Attack**, this chart shows a success rate of 53.2% for the main attack. Despite a majority success, there remains a significant 46.8% failure rate, suggesting some level of resilience in the dataset against this type of inference. **Baseline Attack**, the baseline attack has nearly balanced outcomes with a 49.0% success rate, indicating that about half of the attempts are unsuccessful (51.0% failure rate). This suggests that the basic privacy measures in place provide a near 50-50 chance of protecting the data. And, the **Control Attack** presents a slightly higher success rate at 52.3%, with a 47.7% failure rate. This indicates that this type of attack is slightly more effective, yet still encounters a considerable rate of failures. These charts visually underscore the variable effectiveness of different attack types on the dataset, with none achieving overwhelming success, which speaks to the inherent complexities in fully breaching the synthetic dataset's privacy measures.

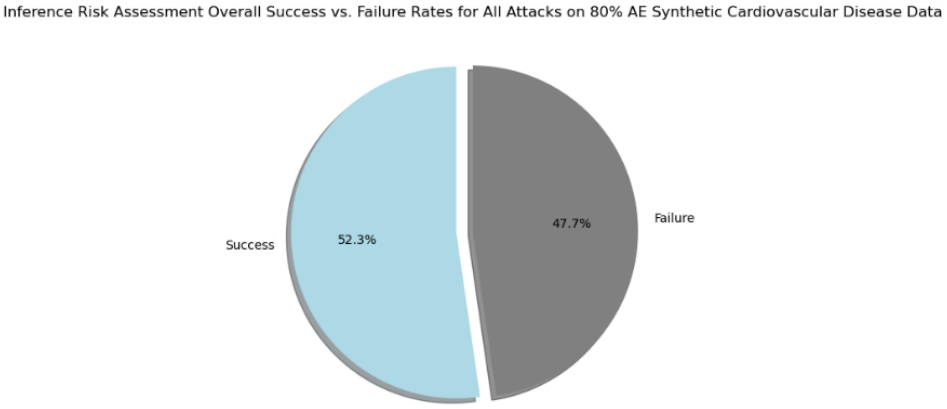


Figure 4.87: Inference Risk Assessments Success/Overall Success vs. Failure Rates by Size of smaller dataset on 80% AE Synthetic Cardiovascular Disease Data

The pie chart in Figure 4.87 illustrates the overall success and failure rates for all attacks combined on the AE Synthetic Cardiovascular Disease Data. It shows that approximately 52.6% of the attacks are successful, while 47.4% are not. This visualization helps in summarizing the collective effectiveness of the attack strategies employed, indicating a relatively balanced outcome between successful and unsuccessful attacks. This balance suggests that while the synthetic data maintains certain vulnerabilities, there is also a significant level of resilience against attempted breaches, reflecting a nuanced security profile of the synthetic dataset.

4.5.5 Evaluation of Privacy Preservation through Singling-Out Univariate Risk Assessment on VAE Synthetic Cardiovascular Data

Table 4.24 presents a comprehensive overview of the univariate singling-out risk assessment for the VAE synthetic cardiovascular data under different attack scenarios. It details success rates for main, baseline, and control attacks, illustrating the minimal variation across different types of evaluations and the number of attacks. The privacy risks remain at zero, with confidence intervals providing a measure of statistical confidence in these estimates. This detailed presentation helps in assessing the robustness of the synthetic data against potential re-identification risks, crucial for its utilization in privacy-sensitive research.

Bar Graph Analysis: The bar graph in Figure 4.88 illustrates the success rates for the singling-out univariate risk assessment at two different scales of attacks—1500 and 500. Notably, the success rates are remarkably low, with a rate of approximately 0.13% for 1500 attacks with grey bars and 0.38% for 500 attacks with green bars. This indicates that the AE synthetic cardiovascular data effectively preserves the privacy of the original data against singling-out

Table 4.24: Detailed Univariate Singling Out Risk Assessment for VAE Synthetic Cardiovascular Data. The table highlights the effectiveness of the synthetic data in preserving privacy against potential re-identification attacks, crucial for its application in privacy-sensitive environments.

	n_attacks=1500	**n_attacks=500**
Main Attack Success Rate	0.0013	0.0038
Baseline Attack Success Rate	0.0013	0.0038
Control Attack Success Rate	0.0013	0.0038
Privacy Risk	0.0	0.0
Confidence Interval	(0.0, 0.00181)	(0.0, 0.00541)

attacks, making it difficult for attackers to identify individual records.

Pie Charts Analysis: The analysis of pie charts for different numbers of attacks provides a comprehensive view of the success versus failure rates in safeguarding privacy. In the scenario with 1500 attacks, the pie chart shows a minimal success rate of 0.13% juxtaposed against a failure rate of 99.87%, which emphasizes the effectiveness of the synthetic data in preventing individual identification. When the number of attacks is reduced to 500, the success rate sees only a slight increase to 0.38%, while the failure rate remains high at 99.62%, further reinforcing the robustness of privacy preservation. Taking into account both scenarios, the overall success rate averages to about 0.25%, underlining consistent protection even under varying levels of adversarial scrutiny.

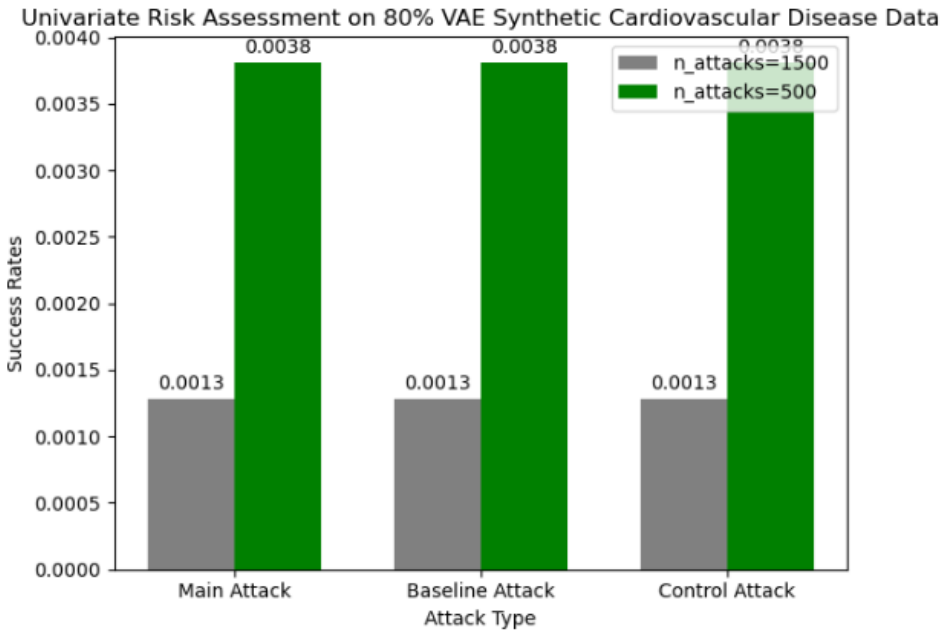


Figure 4.88: Univariate Risk Assessments on 80% VAE Synthetic Cardiovascular Disease Data

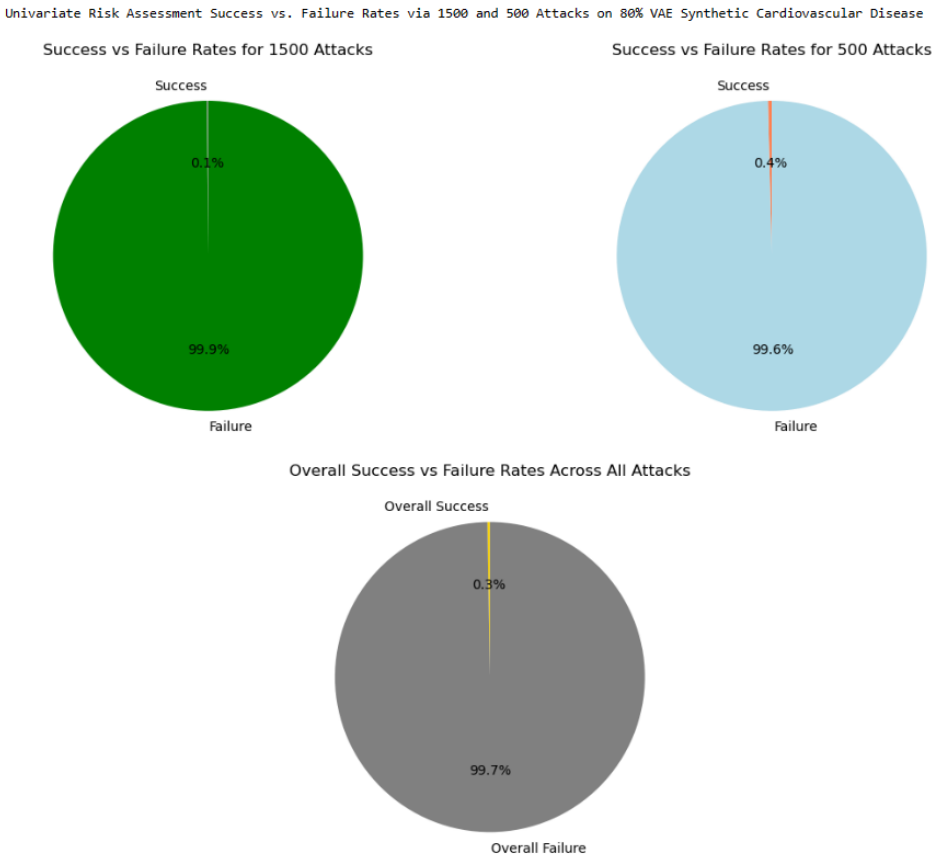


Figure 4.89: Univariate Risk Assessments Success/Overall Success vs. Failure Rates by Number of Attacks on 80% VAE Synthetic Cardiovascular Disease Data

4.5.6 Evaluation of Privacy Preservation through Singling-Out Multivariate Risk Assessment on VAE Synthetic Cardiovascular Data

Table 4.25 encapsulates the outcomes of privacy risk evaluations conducted via multivariate analysis methods. It details the success rates of main, baseline, and control attacks, alongside the privacy risk assessments, across two scenarios: 1500 and 500 attacks. Each entry under the success rates delineates the mean success rate along with the standard error, indicating the consistency and reliability of the synthetic data's ability to protect against privacy breaches. The privacy risk is quantified along with its confidence interval, reflecting the statistical confidence in the data's robustness against identifying individual records. This tabular representation serves as a comprehensive summary to communicate the effectiveness of the synthetic data in safeguarding sensitive information, crucial for stakeholders relying on synthetic datasets for research and development in healthcare.

Table 4.25: Multivariate Singling Out Risk Assessment on VAE Synthetic Cardiovascular Data

Metric/Number of Attacks	1500 Attacks	500 Attacks
Main Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Baseline Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Control Attack Success Rate	0.0013 (±0.0013)	0.0038 (±0.0038)
Privacy Risk	0.0 (CI: 0.0, 0.0018)	0.0 (CI: 0.0, 0.0054)

The bar graph in Figure 4.90 effectively illustrates the differences in success rates between 1500 and 500 attacks, across three distinct types: Main, Baseline, and Control. For each type of attack, there is a noticeable consistency, where the success rates are slightly higher for 500

attacks as opposed to 1500. This trend suggests that the number of attacks may influence the outcomes. The visualization serves as a useful tool in assessing the robustness of the synthetic data model against different intensities of privacy attack simulations.

The pie charts in Figure 4.91 illustrating the success versus failure rates for 1500 and 500 attacks provide a clear visualization of how effective the synthetic data is in thwarting attempts to identify individuals. For the scenario involving 1500 attacks, the success rate stands at a modest 0.39%, demonstrating the robust privacy measures in place. Interestingly, when the number of attacks is reduced to 500, there is an observable increase in the success rate to 1.14%. This change suggests that the synthetic data may exhibit increased vulnerability under less frequent but potentially more focused attacks. Additionally, an overall pie chart that aggregates the outcomes of both attack scenarios offers a comprehensive view of the synthetic data's resilience, with an average success rate of about 0.77%. This rate highlights the general effectiveness of the synthetic data in safeguarding privacy. These visual representations are essential for stakeholders to assess the privacy risks associated with the use of synthetic data. They also serve as a basis for further enhancements to the data generation processes, aiming to improve privacy protections while maintaining the data's practical value.

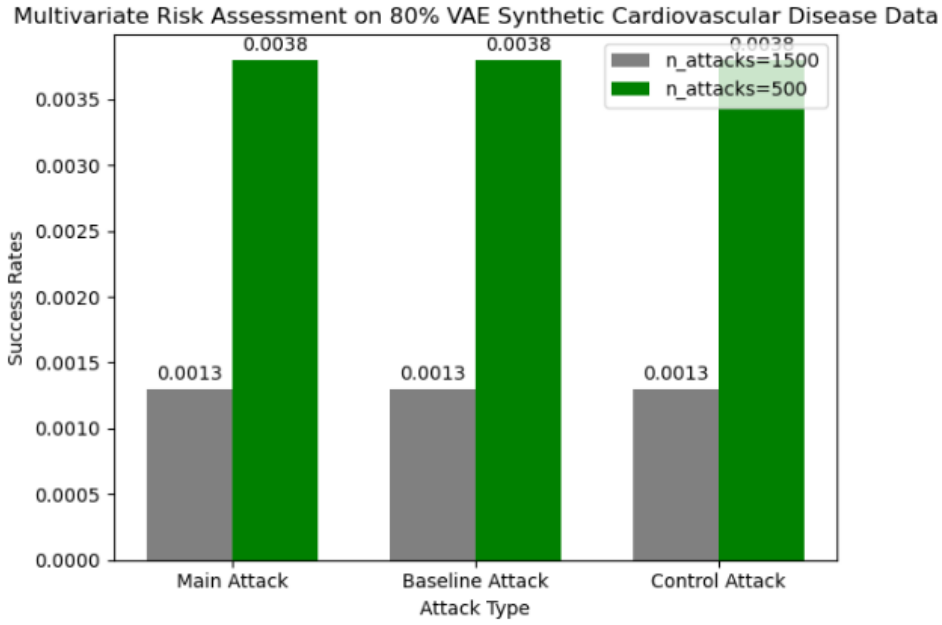


Figure 4.90: Multivariate Risk Assessments on 80% VAE Synthetic Cardiovascular Disease Data

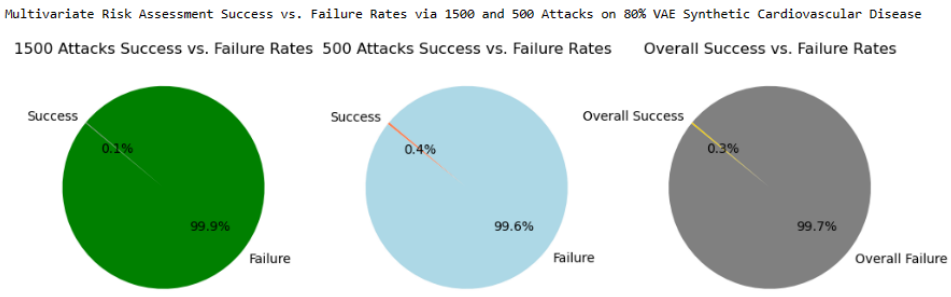


Figure 4.91: Multivariate Risk Assessments Success/Overall Success vs. Failure Rates by Number of Attacks on 80% VAE Synthetic Cardiovascular Disease Data

4.5.7 Evaluation of Privacy Preservation through Linkability Risk Assessment on VAE Synthetic Cardiovascular Data

Table 4.26 shows the linkability risk assessment that was conducted to evaluate the potential of identifying individual records in the AE synthetic dataset that resemble those in the original cardiovascular dataset. Two settings were considered, differing in the number of neighbors used to assess similarity:

Main, Baseline, and Control Attack Success Rate: These rows show the success rates of different types of simulated attacks, quantifying how often the synthetic data could be linked back to individual records in the original dataset. The rates are presented for two settings based on the number of neighbors considered (10 and 5).

Privacy Risk: This row provides a measure of the overall risk to individual privacy, quantifying the likelihood that an individual's data could be re-identified in the synthetic dataset. The confidence intervals (CI) provide a range within which the true privacy risk value is likely to lie, giving an indication of the statistical certainty of the measurements. This structured approach aids in understanding the privacy implications of using synthetic data, enabling stakeholders to make informed decisions about its application in sensitive contexts like healthcare.

Table 4.26: Tabular Representation of Linkability Risk Assessment on VAE Synthetic Cardiovascular Disease Data

Evaluation Metric	**n_neighbors=10**	**n_neighbors=5**
Main Attack Success Rate	0.001065 (±0.000523)	0.000494 (±0.000342)
Baseline Attack Success Rate	0.000851 (±0.000463)	0.000566 (±0.000369)
Control Attack Success Rate	0.000708 (±0.000419)	0.000351 (±0.000279)
Privacy Risk	0.000357 (CI: 0.0, 0.001027);	0.000143 (CI: 0.0, 0.000584)

The bar graph in Figure 4.92 visually compares the success rates of various attack types under two scenarios involving 10 and 5 nearest neighbors, forming part of a linkability risk assessment on VAE Synthetic Cardiovascular Data. This comparison offers several insights into how the model performs under different conditions.

Firstly, the Main Attack shows a slight increase in success rate with more neighbors, implying that the model may be more vulnerable to revealing links when it includes more data points. Conversely, the Baseline Attack, which typically registers lower success rates, exhibits an intriguing rise in success when the number of neighbors is reduced. This observation suggests that the baseline method might be uncovering specific patterns not sufficiently obscured by the synthetic process.

Meanwhile, the success rate for the Control Attack, intended to simulate random guessing, declines with fewer neighbors. This trend suggests that the control setup struggles more in a constrained environment, finding it challenging to predict links accurately.

Each bar in the graph is clearly annotated with the exact success rate, providing a precise numerical understanding of each attack's effectiveness. Such detailed visualization is vital for evaluating the synthetic dataset's resilience against linkability attacks and comprehending how adjustments in evaluation parameters can influence privacy risks.

This pie chart in Figure 4.93 demonstrates the success versus failure rates for attacks using 10 neighbors in linkability risk assessment. The majority of attempts result in failure, underscoring the robustness of the AE synthetic cardiovascular data in resisting this type

of privacy attack. The individual success rates for Main, Baseline, and Control Attacks are relatively low, indicating a strong privacy-preserving characteristic of the synthetic data. **5 and 10 Neighbors Success vs. Failure Rates** With 5 neighbors, the success rates for individual attacks are marginally different from those with 10 neighbors, which again confirms the effectiveness of the data's privacy measures. The slightly higher success rates across attack types do not significantly compromise data privacy, as the vast majority of attacks still result in failure. **Overall Success vs. Failure Rates** The overall pie chart aggregates the success rates from both 10 and 5 neighbors, providing a comprehensive view of the synthetic data's vulnerability to linkability attacks. It reflects a predominant trend towards attack failure, with only 2% overall success, highlighting the synthetic dataset's efficacy in maintaining privacy across different settings and attack complexities. These visualizations serve as a practical tool for understanding the balance between data utility and privacy, suggesting areas for potential improvement in data synthesis processes to better protect sensitive information while maintaining data utility for analytical purposes.

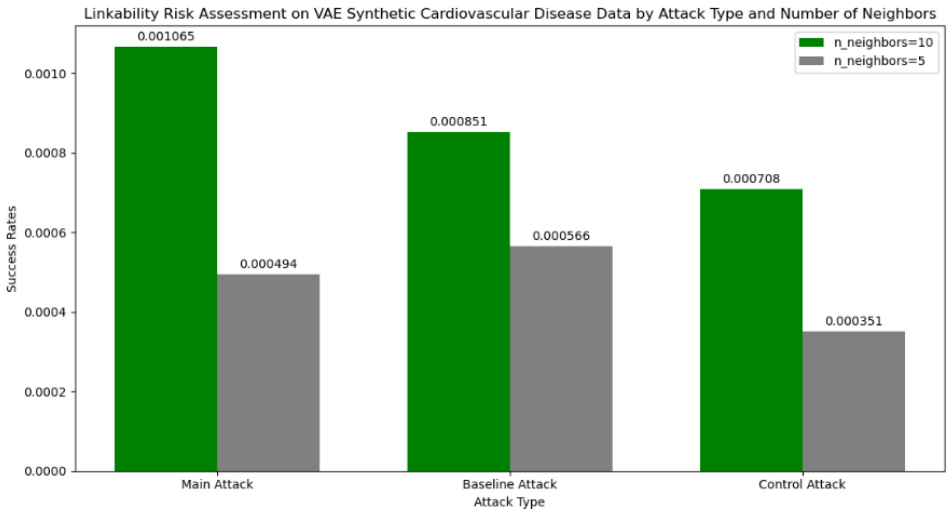


Figure 4.92: Linkability Risk Assessments on 80% VAE Synthetic Cardiovascular Disease Data

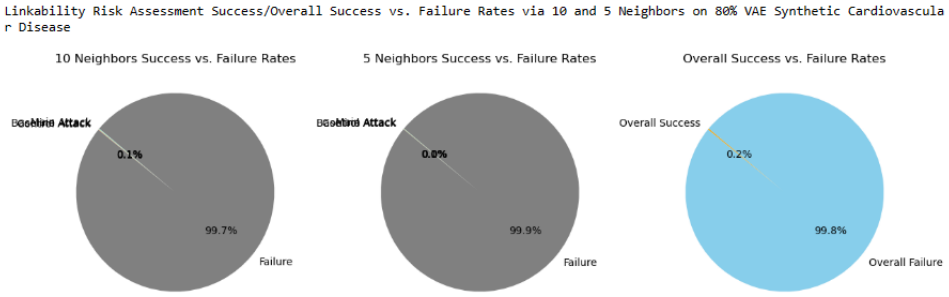


Figure 4.93: Linkability Risk Assessments Success/Overall Success vs. Failure Rates by Number of Attacks on 80% VAE Synthetic Cardiovascular Disease Dataset

4.5.8 Evaluation of Privacy Preservation through Inference Risk Assessment Per-Column on VAE Synthetic Cardiovascular Data

The inference risk assessment for the VAE synthetic cardiovascular data, conducted with the smallest dataset size, highlights a variation in the level of risk across different features. Notably, most features demonstrated minimal risk, with several key features such as **height, aphi, aplo, active, and cardio** exhibiting slightly higher privacy risks. The **main attack** had a

success rate of approximately 60.48%, significantly higher than the baseline (50.47%) but close to the control attack rate (60.01%), suggesting that the synthetic data might not adequately mask the attributes, allowing certain inferences to be made more effectively than random guessing.

Table 4.27 provides a clear overview of which features might be at higher risk of re-identification or information inference, helping guide further privacy enhancements. The risk values are accompanied by confidence intervals and, where relevant, the success rates of different types of inference attacks. This detailed presentation aids in assessing the vulnerability of each attribute in the synthetic dataset, crucial for ensuring the data's utility while safeguarding privacy. This table summarizes the inference risks associated with each attribute in the VAE synthetic cardiovascular data, highlighting both the quantitative privacy risks and the success rates of inference attacks. Such detailed insights are pivotal for assessing the security of synthetic data against potential unauthorized inferences, ensuring its suitability for sensitive applications like healthcare where data privacy is paramount.

Attribute	**Privacy Risk**	**Confidence Interval**	**Success Rate (Main Attack)**
Age	0.0	(0.0, 0.0177)	60.5% (±0.81%)
Gender	0.0	(0.0, 0.0150)	60.5% (±0.81%)
Height	0.3336	(0.1365, 0.5307)	60.5% (±0.81%)
Weight	0.0	(0.0, 0.0182)	60.5% (±0.81%)
Ap_hi	0.0115	(0.0, 0.0257)	60.5% (±0.81%)
Ap_lo	0.0343	(0.0, 0.1089)	60.5% (±0.81%)
Cholesterol	0.0	(0.0, 0.0002)	60.5% (±0.81%)
Gluc	0.0	(0.0, 0.0002)	60.5% (±0.81%)
Smoke	0.0	(0.0, 0.0304)	60.5% (±0.81%)
Alco	0.0	(0.0, 0.0002)	60.5% (±0.81%)
Active	0.0133	(0.0029, 0.0237)	60.5% (±0.81%)
Cardio	0.0116	(0.0, 0.0401)	60.5% (±0.81%)

Table 4.27: Inference Risk and Success Rates for VAE Synthetic Cardiovascular Data. Each row represents a feature with its corresponding estimated privacy risk, confidence interval, and success rates of inference attacks (where applicable). This table provides a clear overview of which features might be at higher risk of re-identification or information inference, helping guide further privacy enhancements.

The bar graph in Figure 4.94 displays the inference risk values for various attributes in the VAE Synthetic Cardiovascular Data, measured when the number of attempted attacks is equal to the size of the smallest dataset used. Attributes like **height** show a notably higher risk value compared to others, indicating specific vulnerabilities in how these features may reveal sensitive information. Most other attributes have minimal to zero risk, suggesting effective anonymization for those characteristics. This visualization is crucial for identifying potential weaknesses in privacy protection and guiding improvements in the synthetic data generation process.
The pie charts in Figure 4.95 illustrate the success versus failure rates for various types of attacks on the VAE Synthetic Cardiovascular Data. The Main Attack, with a success rate of about 60.5%, indicates that this attack method could successfully predict certain elements or patterns in the synthetic data with moderate probability. In contrast, the Baseline Attack shows a lower success rate at 50.5%, suggesting that this simpler or more generic method was less successful, which points to the synthetic data maintaining some resilience against less sophisticated attacks. Interestingly, the Control Attack exhibits a high success rate, nearly 60.0%, almost equal to the Main Attack. This similarity suggests that the control condition, possibly involving different parameters or setups from the main attack, still poses a significant challenge to the privacy preservation mechanisms of the synthetic data.

These visual representations highlight the necessity to balance the utility of synthetic data with robust privacy protection mechanisms, particularly in the face of various adversarial conditions. Each chart clearly marks the proportion of success versus failure, offering an intuitive understanding of how effectively the synthetic data's defenses are holding up against these attacks.

The overall success versus failure rates pie chart in Figure 4.96 for all attack types combined on the VAE Synthetic Cardiovascular Disease Data illustrates a majority of the attacks as successful, with an overall success rate of 56.9% and a failure rate of 43.1%. This indicates a notable vulnerability in the synthetic dataset against these types of privacy attacks. The predominance of successful attacks underscores the need for further enhancements in the methods used to generate synthetic data to ensure better privacy protection. This chart visually represents the proportions of success and failure, highlighting the challenges in protecting privacy in synthetic datasets.

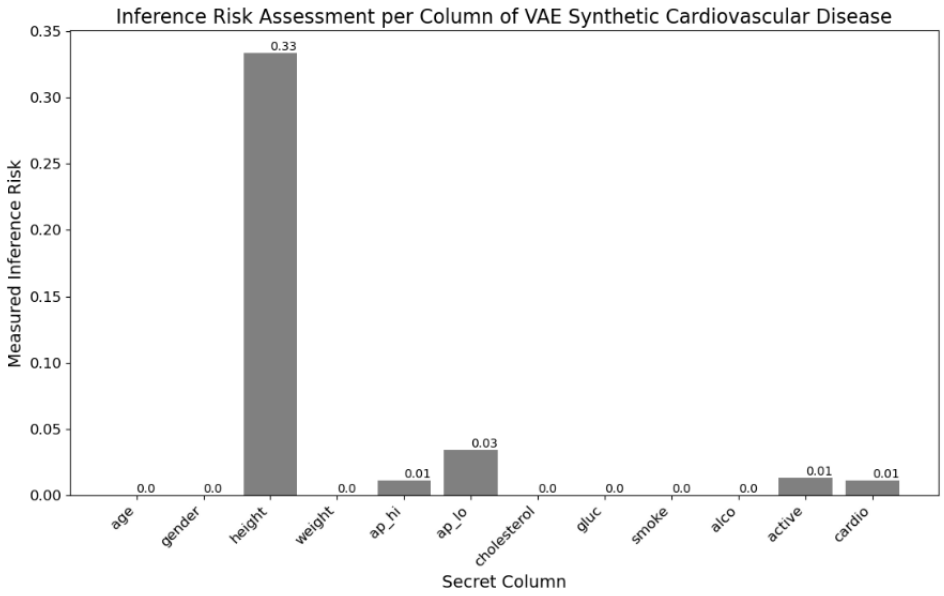


Figure 4.94: Inference Risk Assessments Per Columns on 80% VAE Synthetic Cardiovascular Disease Data

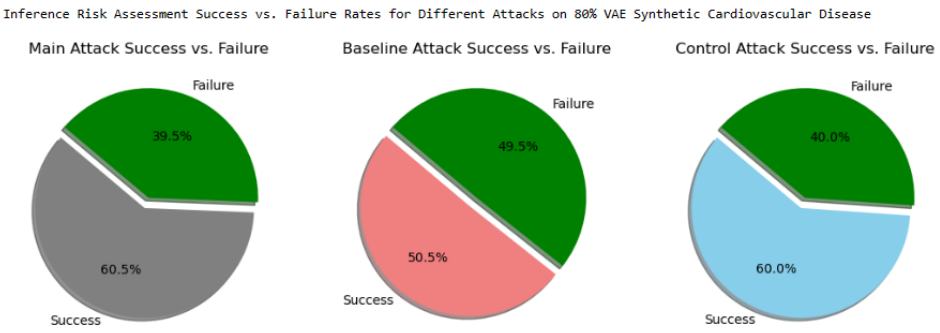


Figure 4.95: Inference Risk Assessments Success/Overall Success vs. Failure Rates by Number of Attacks on 80% VAE Synthetic Cardiovascular Disease Data

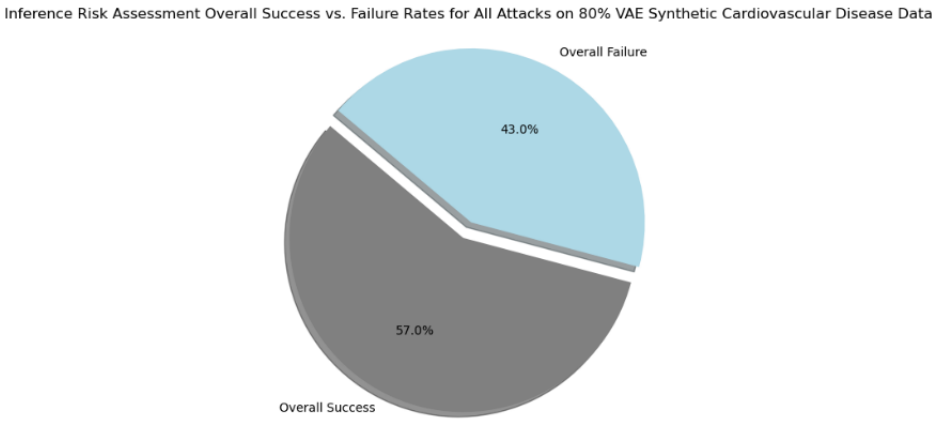


Figure 4.96: Inference Risk Assessments Success/Overall Success vs. Failure Rates by Number of Attacks on 80% VAE Synthetic Cardiovascular Disease Data

4.5.9 Comparative Analysis of Privacy Risk Assessments Between AE-Synthetic and VAE-Synthetic Cardiovascular Disease Data

Singling-Out Risk Assessment

Both the AE and VAE models demonstrate robust privacy preservation capabilities in singling-out risk assessments, as evidenced by Tables 4.20 and 4.24. Both datasets exhibit low success rates for singling-out attacks (Figures 4.79 and 4.88), with nearly zero privacy risk. This similarity underscores that both AE and VAE techniques effectively mitigate the risk of identifying individual records from their respective synthetic datasets.

Multivariate Risk Assessment

The multivariate risk assessments for both datasets (Tables 4.21 and 4.25) also reveal minimal privacy risks, with the AE and VAE models showing equivalent success rates and consistently low privacy risks. This suggests that both models handle multiple attributes with a high degree of privacy, preventing attackers from exploiting combinations of multiple data points.

Linkability Risk Assessment

Linkability assessments (Tables 4.22 and 4.26) show a slight edge in the AE model's ability to prevent linkage between the synthetic data and original datasets. The AE synthetic data exhibits lower success rates in linkability attacks, especially with a lower number of neighbors, suggesting stronger anonymization processes that obscure the relationships between data points more effectively compared to the VAE synthetic data.

Inference Risk Assessment

Inference risk assessments reveal a noticeable difference between the two models. The VAE model exhibits a somewhat higher success rate in inference attacks compared to the AE model, particularly in attributes such as height and blood pressures (Tables 4.23 and 4.27). This indicates that the VAE model may not mask the dependencies and interactions between attributes as effectively as the AE model, potentially allowing more precise inferences about individual data points.

Both AE and VAE synthetic datasets show robust privacy-preserving properties across various assessment metrics. However, slight differences in their performance suggest that the AE model might offer stronger resistance against specific types of privacy attacks such as linkability and certain inference attacks. On the other hand, the VAE model, while generally

effective, shows some vulnerabilities, particularly when fewer attacks are used, which could be an area for improvement in future iterations of synthetic data generation techniques.

This comparative analysis helps underscore the effectiveness of both models in maintaining privacy but also highlights potential areas for enhancing the VAE model's capabilities to match or surpass the privacy standards set by the AE model. This insight is vital for stakeholders considering the use of synthetic data for research and development, especially in fields requiring stringent data privacy measures, such as healthcare.

4.6 Data Utility of Lower Back Pain Data

4.6.1 Comparative Analysis of Original and AE Synthetic Lower Back Pain Data: A Multi-Faceted Evaluation Using AE Model

Chronic lower back pain (CLBP) is recognized as a significant contributor to disability and has seen a dramatic increase in prevalence among adults over the past decade, particularly in older populations. According to Allegri et al. 2016, the incidence of CLBP has more than doubled, highlighting a growing health concern. The complexity of CLBP means that symptoms can vary widely from one individual to another, complicating diagnosis and treatment. Often, the condition requires intricate clinical decision-making that can still lead to misdiagnoses.

Typical sources of lower back pain include irritation of the large nerve roots that extend to the legs, irritation of the smaller nerves that serve the lower back, strain in the large paired lower back muscles known as the erector spinae, damage to the bones, ligaments, or joints, degeneration of an intervertebral disc, and issues with any of these structures can lead to lower back pain that may radiate or refer to other body parts. Furthermore, many lower back problems are also associated with muscle spasms, which, though they may seem minor, can lead to severe pain and significant disability.

The analysis demonstrates varying degrees of similarity between the original and synthetic datasets across different features. Notably, the pelvicradius and degreespondylolisthesis features showed the most significant differences in distributions (KS-Test P-Values < 0.01). Error metrics such as MSE, RMSE, and MAE were particularly high for degreespondylolisthesis, indicating substantial differences in individual data points between the datasets. Table 4.28 encapsulates the results from statistical tests assessing the differences in data distributions between the original and AE synthetic datasets. It features results from the Kolmogorov-Smirnov test (KS Test), F-Test, and T-Test, which are crucial for understanding the statistical similarities or discrepancies between the two data sets. Table 4.29 complements the statistical tests by presenting detailed error metrics and basic statistical comparisons such as mean and standard deviation for each feature. This provides a comprehensive view of the error dynamics and central tendencies which are vital for assessing the quality and usability of synthetic data in practical applications.

Table 4.28: Statistical Tests for Original and AE Synthetic Lower Backpain Data

Feature	KS Stat	KS P-Value	F Stat	F P-Value	T Stat	T P-Value
Pelvic Incidence	0.089	0.284	1.750	0.187	1.323	0.187
Pelvic Tilt	0.109	0.106	0.044	0.833	0.211	0.833
Lumbar Lordosis Angle	0.105	0.131	1.637	0.201	1.280	0.201
Sacral Slope	0.125	0.041	3.373	0.067	1.837	0.067
Pelvic Radius	0.246	0.000005	0.101	0.750	0.318	0.750
Degree Spondylolisthesis	0.153	0.006	4.295	0.039	2.072	0.039

Table 4.29: Error Metrics and Basic Statistics for Original and AE Synthetic Lower Backpain Data

Feature	MSE	RMSE	MAE	Mean (Orig, Syn)	Std (Orig, Syn)
Pelvic Incidence	52.602	7.253	5.594	(59.915, 57.999)	(17.037, 15.172)
Pelvic Tilt	54.818	7.404	5.773	(16.882, 16.730)	(9.519, 6.120)
Lumbar Lordosis Angle	106.820	10.335	7.713	(51.236, 49.403)	(17.814, 13.834)
Sacral Slope	76.842	8.766	6.861	(43.033, 41.061)	(13.259, 10.498)
Pelvic Radius	143.503	11.979	8.992	(117.936, 117.666)	(12.643, 4.313)
Degree Spondylolisthesis	568.285	23.839	15.263	(25.362, 19.107)	(38.262, 28.201)

Graphical Interpretations and Statistical Insights Between Original and AE-Synthetic Lower Back Pain Data

The bar graph in Figure 4.97 presents the P-values from Kolmogorov-Smirnov (KS), F-Test, and T-Test for various features of the lower back pain dataset, comparing the original data with the AE synthetic data. The KS-Test, which measures the largest difference between the empirical distribution functions of the two samples, shows that P-values closer to 1 suggest similarity. In contrast, values closer to 0 indicate significant differences. Notably, features such as **Pelvic Radius and Degree Spondylolisthesis** exhibit significant differences with P-values less than 0.01, highlighting discrepancies between the original and synthetic distributions. The F-Test assesses the equality of variances between the two samples, where features like **Degree Spondylolisthesis** demonstrate significant differences in variances with P-values less than 0.05. Similarly, the T-Test, which tests the hypothesis that the means of two populations are equal, observes significant P-values below 0.05 in similar features, suggesting noticeable differences in mean values between the datasets.

The scatter plot in Figure 4.98 compares the F-Test and T-Test P-Values for various features from the 80% Original and AE Synthetic Lower Backpain data, where each point represents a feature plotted with its F-Test P-Value on the x-axis and T-Test P-Value on the y-axis. The red dashed line indicates perfect agreement, signifying that if a point lies on this line, both tests yielded similar statistical significance levels for that feature. Most features demonstrate a high degree of alignment between the F-Test and T-Test P-Values, as evidenced by the proximity of points to the line of perfect agreement. This alignment suggests consistent results across both types of tests concerning the statistical differences between the original and synthetic datasets for those features. Notably, features like **Degree Spondylolisthesis and Sacral Slope** exhibit significantly lower P-Values (below 0.1), indicating significant differences in variance and means between the datasets for these features. This graph plays a crucial role in assessing the statistical equivalence of the original and synthetic datasets, offering insights into which features may require further adjustment in the synthetic data generation process to better match the original data's statistical properties.

The graph in Figure 4.99 illustrates the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) for each feature. MSE and RMSE highlight the average squared differences and the square root of these differences, respectively, showing how closely the AE synthetic data approximates the original data. Notably, the **Degree Spondylolisthesis** feature shows the highest errors, indicating considerable discrepancies in this feature between the original and synthetic datasets. The MAE provides a direct average of absolute errors, which is particularly high for **Degree Spondylolisthesis**, further confirming its significant deviation in the synthetic dataset compared to the original.

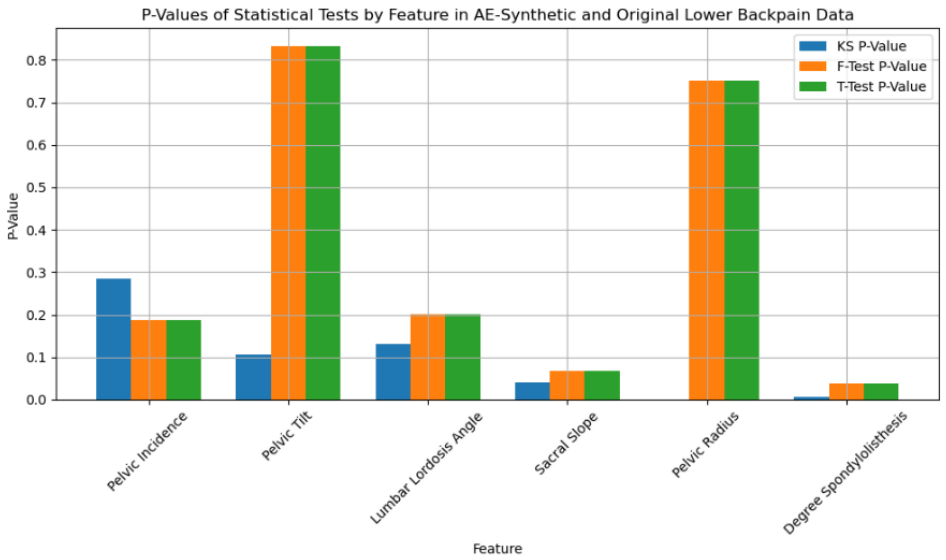


Figure 4.97: P-Values of Statistical Tests by Feature in AE-Synthetic and Original Lower Back Pain Data

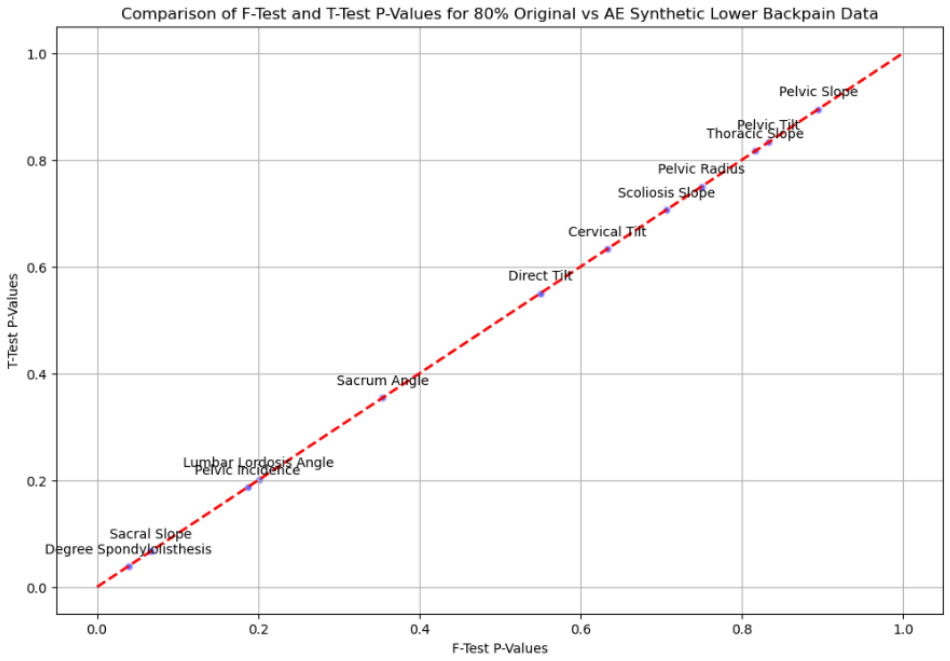
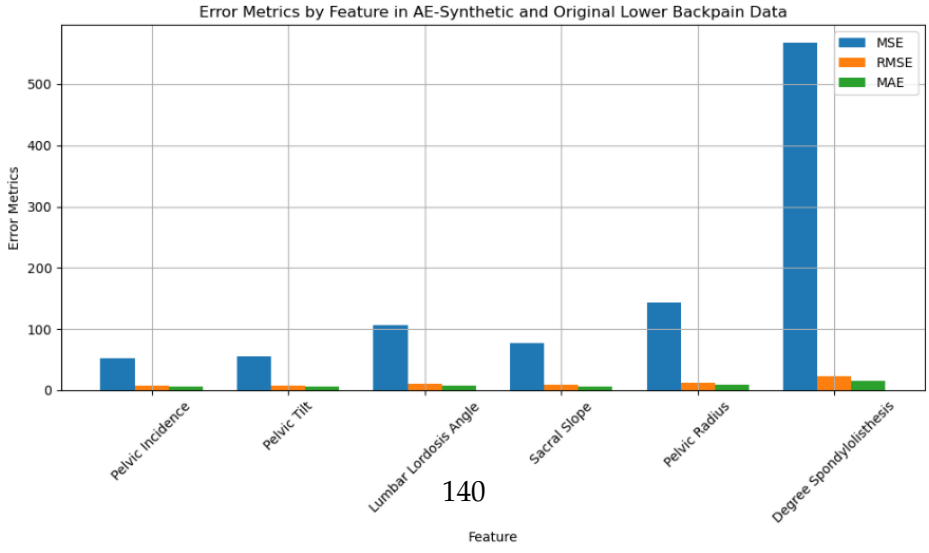


Figure 4.98: Comparison of F-Test and T-Test P-Values for 80% Original and AE-Synthetic Lower Back Pain Data



Correlation Matrix Comparison Between Original and AE-Synthetic Lower Back Pain Data

The graphical correlation matrices of the 80% Original Lower Backpain and 80% AE-Synthetic Lower Backpain in Figure 4.119 and Figure 4.101 reveals interesting insights into how relationships between features are represented in both datasets:

In the Original Lower Backpain Data, pelvic incidence is notably linked with pelvic tilt (0.63), lumbar lordosis angle (0.73), and sacral slope (0.83), reflecting strong connections among crucial spinal alignment metrics. Degree spondylolisthesis also shows moderate associations with lumbar lordosis angle (0.50), sacral slope (0.57), pelvic tilt (0.36), and pelvic incidence (0.65), indicating significant relationships with spinal curvature and alignment. Lumbar lordosis angle's correlation with pelvic tilt (0.42) and sacral slope (0.63) are further observed alongside negative correlations between sacral slope and pelvic radius (-0.34), and pelvic incidence and pelvic radius (-0.27). Additionally, a negative correlation is evident between lumbar lordosis angle and direct tilt (-0.13), highlighting lesser-known interconnections. Figure 4.119

Conversely, the AE Synthetic Lower Backpain Data exhibits generally stronger correlations. Pelvic incidence demonstrates very high correlations with pelvic tilt (0.86), lumbar lordosis angle (0.97), and sacral slope (0.95). Degree spondylolisthesis in the synthetic data shows pronounced connections with lumbar lordosis angle (0.86), sacral slope (0.90), pelvic tilt (0.67), and pelvic incidence (0.89), exceeding the original dataset's values. Enhanced correlations in synthetic data include lumbar lordosis angle with pelvic tilt (0.80) and sacral slope (0.95). Notably, more significant negative correlations such as between sacral slope and pelvic radius (-0.72), and pelvic incidence and pelvic radius (-0.79) are observed, with a newly introduced positive correlation between pelvic radius and direct tilt (0.54). Figure 4.101

This analysis underscores that while the AE-Synthetic dataset amplifies and sometimes introduces new correlations, impacting its realism and potential clinical applicability, it also emphasizes the need for thorough evaluation when utilizing synthetic datasets in research and clinical practices, particularly considering their altered correlation structures compared to original data.

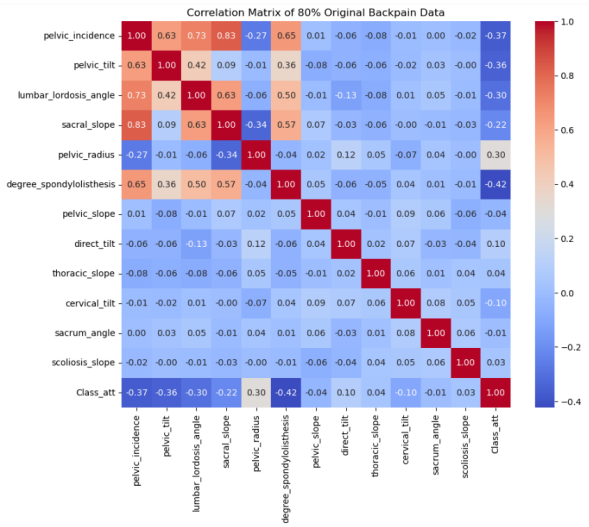


Figure 4.100: Original Lower-Backpain Data

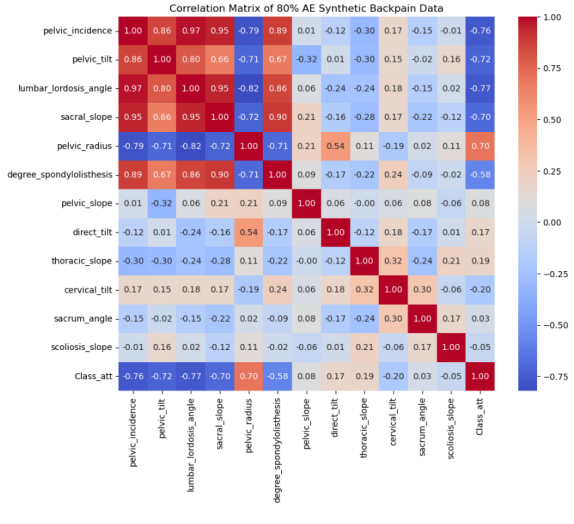


Figure 4.101: AE Synthetic Lower-Backpain Data

The evaluation of the correlation matrices for the 80% Original Lower Backpain and AE Synthetic Lower Backpain data highlights significant insights into the relational dynamics of

various physical measurements associated with lower back pain. In the original dataset, the correlations suggest moderate to strong relationships among pelvic incidence, lumbar lordosis angle, and sacral slope, with pelvic incidence showing a particularly strong relationship with sacral slope (r=0.831). Conversely, correlations with pelvic radius and degree of spondylolisthesis indicate inverse or weak relationships. Figure 4.102

The AE synthetic data mirrors these relationships but with generally enhanced correlation coefficients, implying a more pronounced interdependence among features. Notably, the lumbar lordosis angle and sacral slope exhibit a stronger correlation in synthetic data (r=0.947) compared to the original (r=0.629). This enhancement could suggest that the synthetic generation process amplifies underlying patterns present in the original data. Figure 4.102

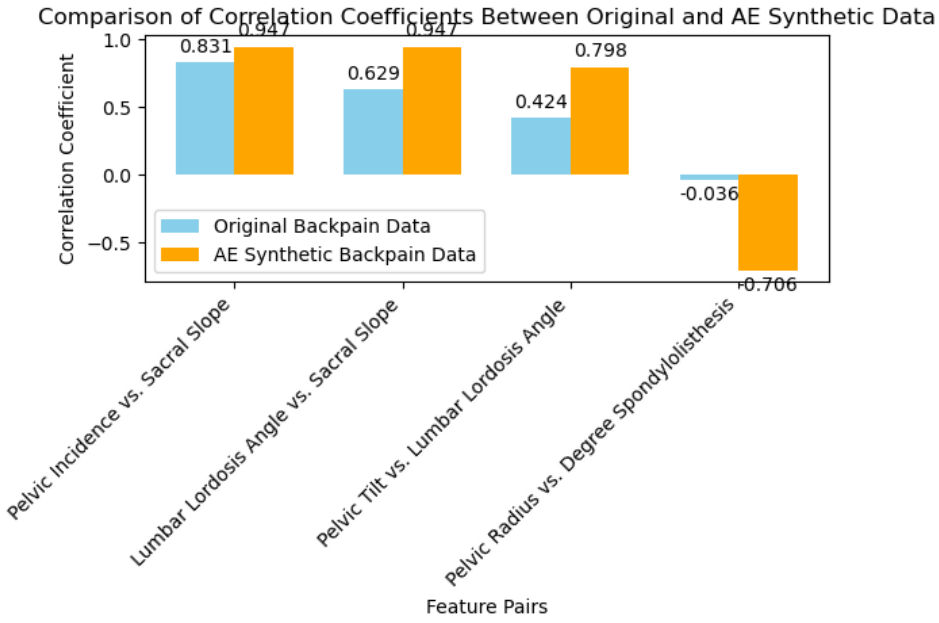


Figure 4.102: Comparison of Correlation Coefficients Between AE-Synthetic and Original Lower Back Pain Data

Area Under Curve Scores By Classifier Between Original and AE-Synthetic Lower Back Pain Data

The bar graph in Figure 4.105 displays a comparison of AUC scores for various classifiers applied to both AE Synthetic and Original Lower Back Pain data. Each classifier's performance is quantified by the AUC (Area Under the Curve) metric, which ranges from 0 to 1, where a higher score indicates better classification performance. The colors, brown for AE Synthetic and yellow for Original, help visually segregate the two types of data for direct comparison. Figure 4.104 and Figure 4.103

In this analysis, the AUC scores reveal a mixed pattern of performance across the two data types. The Random Forest classifier shows equal proficiency on both datasets with an AUC of 0.80, suggesting robustness irrespective of data origin. Similarly, the Gradient Boosting classifier maintains a consistent AUC of 0.78 across both data types.
The Logistic Regression and MLP Classifiers particularly excel with AE Synthetic data, achieving AUCs of 0.90 and 0.90 respectively, compared to 0.82 and 0.80 on the Original data. This indicates a better alignment or perhaps an overfitting to the characteristics of the synthetic data.
The SVC Classifier demonstrates the most pronounced improvement in the synthetic dataset, with an AUC of 0.95 compared to 0.85 on the Original data, suggesting that the characteristics

of the synthetic data might be better captured by this model's learning algorithm.
Conversely, the XGB and LGBM classifiers exhibit lower performance on the synthetic data with AUCs of 0.68 and 0.70, significantly underperforming compared to their original data scores of 0.80 and 0.82. This could indicate challenges in generalizing the learning from synthetic to real-world data or vice versa.
The KNN and Adaboost Classifiers show varied but generally high performance, with the KNN Classifier scoring higher on the synthetic data (AUC of 0.93) compared to the original (AUC of 0.88), and the Adaboost Classifier maintaining a solid performance with AUCs of 0.80 and 0.78 respectively.

Clinical Relevance: The higher AUC scores in synthetic data across most classifiers are promising for model training and preliminary testing. However, the discrepancies observed between synthetic and original data performances necessitate careful validation and potentially calibration when applying these models in real-world clinical settings to ensure reliable performance when faced with actual patient data.

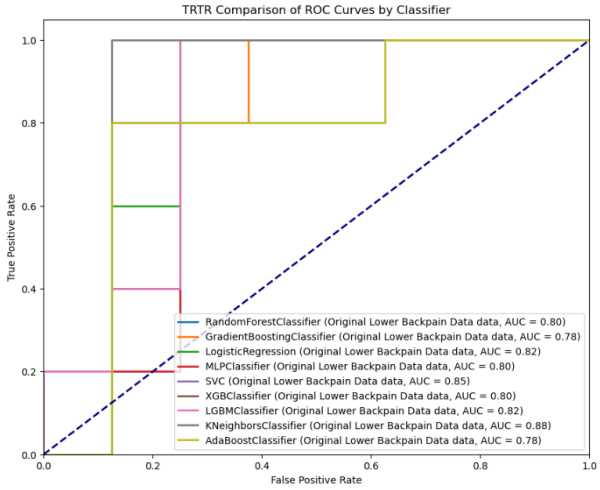


Figure 4.103: AUC-ROC Curve for TRTR

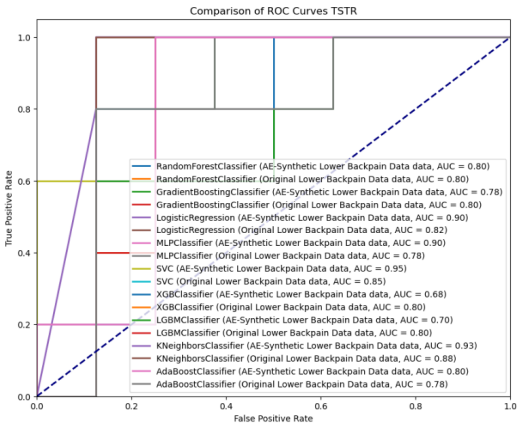


Figure 4.104: Comparison of ROC Curves Original and AE-Synthetic (TRTR and TSTR)

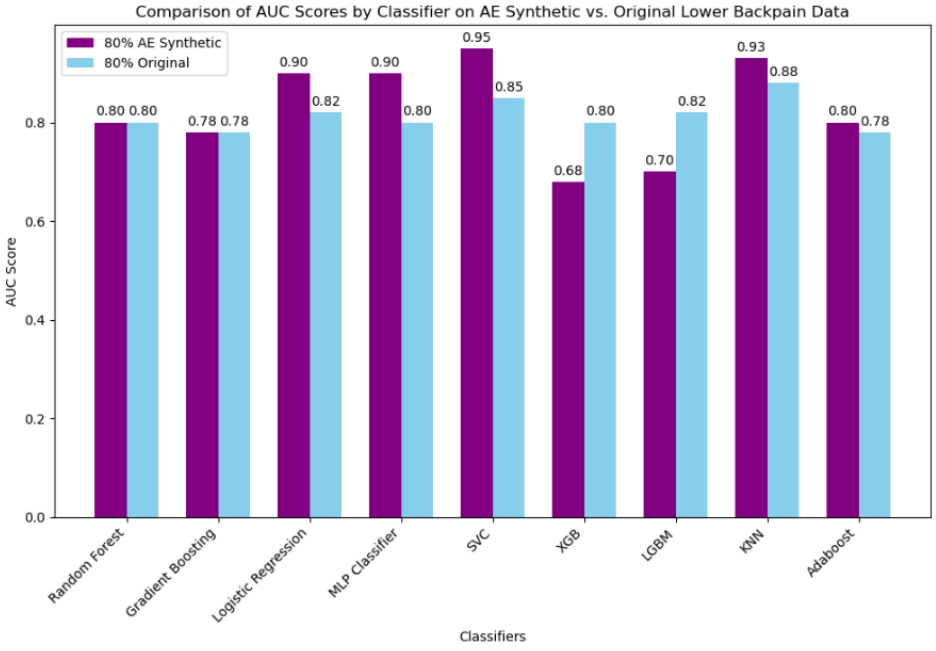


Figure 4.105: Area Under Curve Scores By Classifier Between 80% Original and AE-Synthetic Lower Back Pain Data - Back downward to Table 4.42

Mean Cross-Validation Accuracy By Classifier Between Original and AE-Synthetic Lower Back Pain Data

The analysis of the Mean Cross-Validation (CV) Accuracy between the AE-Synthetic and Original Lower Backpain datasets provides significant insights into the efficacy and reliability of synthetic data in training machine learning models. This comparison highlights the synthetic data's capability to not only mimic but in some cases, surpass the performance metrics of the original dataset.

Starting with Decision Tree classifiers, the synthetic dataset shows an impressive mean CV accuracy of 0.9391 compared to 0.7883 on the original dataset, demonstrating a substantial improvement in model performance when trained on synthetic data. This trend is similarly observed in other classifiers. For instance, the Gradient Boosting and RandomForest classifiers exhibit nearly identical improvements in accuracy on synthetic data, with mean CV accuracies slightly above 93.9%, a significant increase from their performance on the original data which hovers around 82% to 83.8%. More details are seen in Figure 4.106.

The AdaBoostClassifier, while still showing improved performance on synthetic data at 92.9% accuracy, indicates a narrower margin compared to its performance on the original dataset, which achieves an accuracy of 84.37%. This suggests that while synthetic data enhances model performance, the extent of improvement can vary depending on the algorithm's sensitivity to data variations. More details are seen in Figure 4.106.

Similarly, both LGBM and XGB classifiers perform better on the synthetic dataset with accuracies just over 94.4%, compared to around 82.35% and 83.36% on the original data. This consistency across tree-based models indicates a robustness in the AE synthetic data that supports complex decision boundaries effectively. More details are seen in Figure 4.106.

Interestingly, the MLPClassifier shows the most dramatic improvement, with a mean CV accuracy of 95.44% on synthetic data compared to only 64.15% on the original dataset. This stark difference could be attributed to the synthetic data's ability to provide a more regularized or uniform feature representation that benefits the neural network's learning process. More details are seen in Figure 4.106.

However, for simpler models like the KNeighborsClassifier and Logistic Regression, while there is an improvement in performance on synthetic data, it is less pronounced. This might indicate that these models, which rely heavily on the underlying distribution of data, are less capable of leveraging the modifications inherent in synthetic data for substantial performance gains. More details are seen in Figure 4.106.

In the Support Vector Classifier, the performance on synthetic data, with an accuracy of 92.42%, is notably higher than the 82.35% observed with the original data. This reinforces the notion that synthetic data, by potentially smoothing out noise and outliers, might be more amenable to the margin optimization inherent in SVCs. More details are seen in Figure 4.106.

Overall, the consistent outperformance of classifiers on synthetic data underscores its potential utility in scenarios where privacy concerns or data availability limit the use of real data. The enhanced generalization and robustness indicated by higher accuracy scores suggest that synthetic data generation techniques, particularly those utilizing autoencoders as in this study, are advancing towards producing highly reliable and usable datasets. More details are seen in Figure 4.106.

This analysis not only affirms the value of synthetic data in training predictive models but also prompts a deeper discussion on the ethical implications and practical applications of such data in clinical settings. The findings advocate for further research into optimizing synthetic data generation processes to achieve even higher fidelity and performance across a broader array

of machine learning models and scenarios.

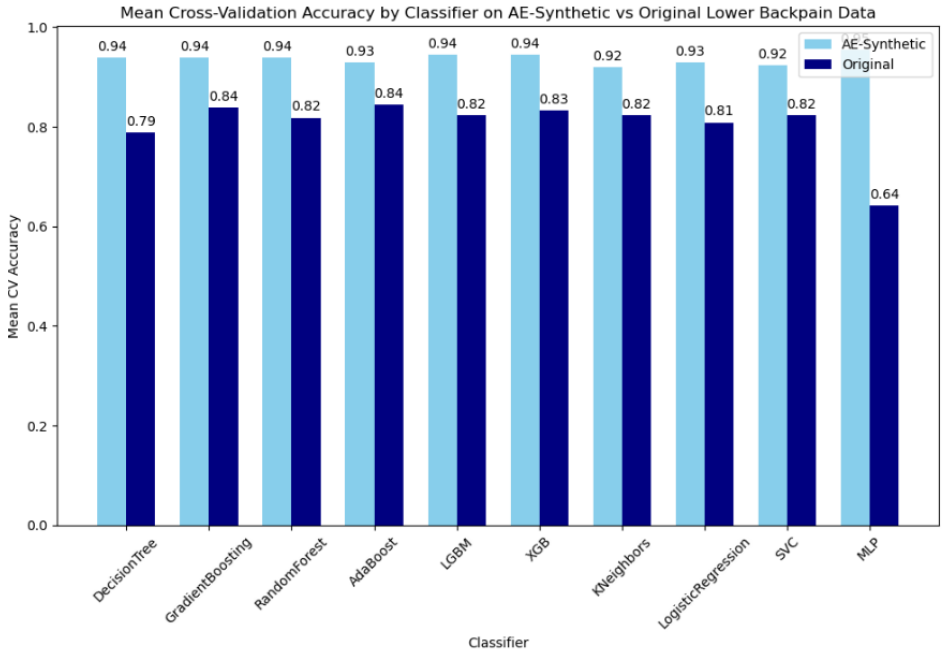


Figure 4.106: Mean Cross-Validation Accuracy By Classifier Between 80% Original and AE-Synthetic Lower Back Pain Data - Back downward to Table 4.42

Classification Reports By Classifier Between Original and AE-Synthetic Lower Back Pain Data

The analysis of different classifiers on the original and AE-Synthetic Lower Backpain data tested against 20% control data reveals varying degrees of effectiveness in preserving the predictive characteristics after synthetic data generation. The blue bars represent the accuracy scores on the original dataset, while the sky blue bars represent scores on the ae-synthetic dataset. Figure 4.115

Gradient Boosting (GB) showed a notable disparity in performance between the original and synthetic datasets. The original data achieved an accuracy of 69%, with a skewed precision favoring negative cases (class 0), whereas the synthetic data mirrored this accuracy but exhibited a more balanced precision between classes. This difference underscores the impact of synthetic data on model behavior and classification balance. Figure 4.107

The Random Forest (RF) classifiers demonstrated an increase in overall accuracy from 62% with the original data to 77% with the synthetic data, suggesting an enhanced generalization capability of the model when trained on synthetic data. This enhancement indicates that synthetic data may help in reducing overfitting and improving model robustness. Figure 4.108

XGBoost (XGB) results were consistent across both data types, with each setup showing an accuracy of around 69%. This consistency indicates that the synthetic data preserved essential data characteristics effectively, maintaining performance levels comparable to the original data. Figure 4.111

Multi-Layer Perceptron (MLP) and Decision Tree Classifier (DCT) both reported higher accuracy on synthetic data (85%) compared to their performance on original data (69% and 77% respectively). This improvement suggests that synthetic data may be providing a clearer or more definitive feature set for classification, potentially due to reduced noise or more distinct feature delineations. Figure 4.110

Logistic Regression (LGR) and K-Nearest Neighbors (KNN) displayed similar trends, with slight improvements in classification metrics on synthetic data, hinting at subtle enhancements in data quality or representation that benefit these models. Figure 4.113 and Figure 4.114

LightGBM (LGBM) and Support Vector Classifier (SVC) showed no significant difference in performance, maintaining consistent accuracy levels across both data types. This outcome indicates that for some models, the synthetic data replicates the statistical properties of the original data closely enough to achieve similar predictive outcomes. Figure 4.112 and Figure 4.109

These findings collectively emphasize the potential of synthetic datasets to serve as viable alternatives or supplements to real datasets in training machine learning models, particularly in scenarios where data privacy is paramount or where original data is limited or biased. However, the variability in performance across different classifiers and metrics suggests that careful consideration is needed when selecting and tuning models for tasks involving synthetic data.

GB on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.67	1.00	0.80	8
1	1.00	0.20	0.33	5
accuracy			0.69	13
macro avg	0.83	0.60	0.57	13
weighted avg	0.79	0.69	0.62	13

GB on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.60	0.60	0.60	5
accuracy			0.69	13
macro avg	0.68	0.68	0.68	13
weighted avg	0.69	0.69	0.69	13

Figure 4.107: GB Classification Report

RF on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.64	0.88	0.74	8
1	0.50	0.20	0.29	5
accuracy			0.62	13
macro avg	0.57	0.54	0.51	13
weighted avg	0.58	0.62	0.56	13

RF on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.86	0.75	0.80	8
1	0.67	0.80	0.73	5
accuracy			0.77	13
macro avg	0.76	0.78	0.76	13
weighted avg	0.78	0.77	0.77	13

Figure 4.108: RF Classification Report

SVC on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.67	0.75	0.71	8
1	0.50	0.40	0.44	5
accuracy			0.62	13
macro avg	0.58	0.57	0.58	13
weighted avg	0.60	0.62	0.61	13

SVC on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	1.00	0.62	0.77	8
1	0.62	1.00	0.77	5
accuracy			0.77	13
macro avg	0.81	0.81	0.77	13
weighted avg	0.86	0.77	0.77	13

Figure 4.109: SVC Classification Report

MLP on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.78	0.88	0.82	8
1	0.75	0.60	0.67	5
accuracy			0.77	13
macro avg	0.76	0.74	0.75	13
weighted avg	0.77	0.77	0.76	13

MLP on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	1.00	0.75	0.86	8
1	0.71	1.00	0.83	5
accuracy			0.85	13
macro avg	0.86	0.88	0.85	13
weighted avg	0.89	0.85	0.85	13

Figure 4.110: MLP Classification Report

XGB on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.67	0.75	0.71	8
1	0.50	0.40	0.44	5
accuracy			0.62	13
macro avg	0.58	0.57	0.58	13
weighted avg	0.60	0.62	0.61	13

XGB on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.60	0.60	0.60	5
accuracy			0.69	13
macro avg	0.68	0.68	0.68	13
weighted avg	0.69	0.69	0.69	13

Figure 4.111: XGB Classification Report

LGBM on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.70	0.88	0.78	8
1	0.67	0.40	0.50	5
accuracy			0.69	13
macro avg	0.68	0.64	0.64	13
weighted avg	0.69	0.69	0.67	13

LGBM on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.86	0.75	0.80	8
1	0.67	0.80	0.73	5
accuracy			0.77	13
macro avg	0.76	0.78	0.76	13
weighted avg	0.78	0.77	0.77	13

Figure 4.112: LGBM Classification Report

LGR on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.60	0.60	0.60	5
accuracy			0.69	13
macro avg	0.68	0.68	0.68	13
weighted avg	0.69	0.69	0.69	13

LGR on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	0.86	0.75	0.80	8
1	0.67	0.80	0.73	5
accuracy			0.77	13
macro avg	0.76	0.78	0.76	13
weighted avg	0.78	0.77	0.77	13

Figure 4.113: LGR Classification Report

KNN on 80% Original BackPain/Tested on 20% Control Data (TRTR):

	precision	recall	f1-score	support
0	0.86	0.75	0.80	8
1	0.67	0.80	0.73	5
accuracy			0.77	13
macro avg	0.76	0.78	0.76	13
weighted avg	0.78	0.77	0.77	13

KNN on 80% AE Synthetic BackPain/Tested on 20% Control Data (TSTR):

	precision	recall	f1-score	support
0	1.00	0.62	0.77	8
1	0.62	1.00	0.77	5
accuracy			0.77	13
macro avg	0.81	0.81	0.77	13
weighted avg	0.86	0.77	0.77	13

Figure 4.114: KNN Classification Report

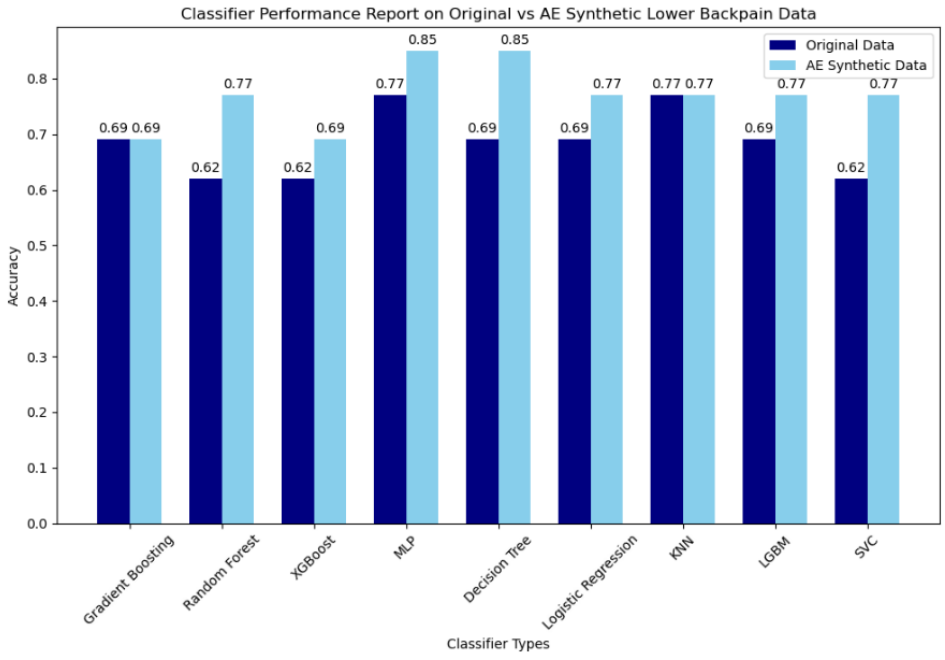


Figure 4.115: Classification Reports By Classifier Between 80% Original and AE-Synthetic Lower Back Pain Data - Back downward to Table 4.42

4.6.2 Comparative Analysis of Original and VAE Synthetic Lower Back Pain Data: A Multi-Faceted Evaluation Using VAE Model

Table 4.30: Statistical Tests for Original and VAE Synthetic Lower Backpain Data

Feature	KS Stat	KS P-Value	F Stat	F P-Value	T Stat	T P-Value
Pelvic Incidence	0.343	2.60×10^{-13}	5.117	0.0241	2.262	0.0241
Pelvic Tilt	0.274	1.31×10^{-8}	1.504	0.2207	1.226	0.2207
Lumbar Lordosis Angle	0.371	1.42×10^{-15}	7.256	0.0073	2.694	0.0073
Sacral Slope	0.395	1.13×10^{-17}	4.694	0.0307	2.167	0.0307
Pelvic Radius	0.363	6.58×10^{-15}	2.428	0.1199	-1.558	0.1199
Degree Spondylolisthesis	0.222	9.34×10^{-6}	5.605	0.0183	2.367	0.0183

Table 4.31: Error Metrics and Basic Statistics for Original and AE Synthetic Lower Backpain Data

Feature	MSE	RMSE	MAE	Mean (Orig, Syn)	Std (Orig, Syn)
Pelvic Incidence	359.487	18.960	15.121	(59.915, 57.232)	(17.037, 7.646)
Pelvic Tilt	110.349	10.505	8.213	(16.882, 16.068)	(9.519, 4.320)
Lumbar Lordosis Angle	363.977	19.078	15.019	(51.236, 48.012)	(17.814, 6.153)
Sacral Slope	192.003	13.857	10.858	(43.033, 41.150)	(13.259, 3.389)
Pelvic Radius	179.772	13.408	10.078	(117.936, 119.244)	(12.643, 3.879)
Degree Spondylolisthesis	1820.928	42.672	26.587	(25.362, 19.071)	(38.262, 16.938)

The Statistical Tests Table in Table 4.30 showcases the results of the KS-Test, F-Test, and T-Test for each feature, effectively highlighting significant differences in distributions, variances, and means between the AE Synthetic and Original datasets. In contrast, the Error Metrics Table in Table 4.31 provides a detailed account of the Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and fundamental statistics like mean and standard deviation for each feature, illustrating the levels of error and variation observed within the datasets. Collectively, these tables offer a comprehensive and succinct overview of the statistical discrepancies and the magnitudes of errors encountered when analyzing features across the two datasets, which is crucial for assessing the quality of the datasets and the performance of models developed from them.

Graphical Interpretations and Statistical Insights Between Original and VAE-Synthetic Lower Back Pain Data

The visual representations above illustrate the P-values of various statistical tests and the error metrics (MSE, RMSE, MAE) for different spinal parameters, comparing AE-Synthetic and Original Lower Backpain Data. These graphs provide a comprehensive view of how these metrics diverge across the two types of datasets.

P-Values Graph

- **KS P-Values:** Displaying the smallest P-values, highlighting significant differences in distributions between the AE-Synthetic and Original datasets for all features. The logarithmic scale emphasizes extremely small values, indicating strong statistical evidence against the null hypothesis of identical distributions.
- **F-Test and T-Test P-Values:** These bars reflect the tests for equality of variances and means, respectively. Except for a few cases where the F-test shows significance (like for pelvic incidence and lumbar lordosis angle), most P-values are not below the traditional

alpha level of 0.05, indicating insufficient evidence to reject the null hypothesis of equal variances and means for several features.

Error Metrics Graph

- **Mean Square Error (MSE):** Displays substantial variability across features, with the degree of spondylolisthesis showing the highest error, suggesting significant challenges in predictive accuracy for this parameter.

- **Root Mean Square Error (RMSE) and Mean Absolute Error (MAE):** These metrics track closely with MSE and provide a sense of the average magnitude of errors. The patterns indicate that while some features like pelvic radius are predicted with relatively lower error, others like degree of spondylolisthesis exhibit high error rates, underscoring potential issues in model performance or data complexity.

Clinical Relevance: These analyses are crucial for understanding model reliability and the appropriateness of synthetic data for training predictive models. The significant statistical differences highlighted by the **KS-tests** necessitate careful consideration and validation of models intended for clinical use, ensuring they perform adequately when confronted with real, varied clinical data. The error metrics provide additional insight into potential practical limitations of current modeling approaches, particularly in handling complex features such as degree spondylolisthesis. This understanding can guide further model refinement and data collection strategies to enhance the robustness and accuracy of predictive models in clinical settings. The close alignment of **F-Test and T-Test P-values** across most features indicates that the variability and mean differences between AE-Synthetic and Original datasets are consistently recognized by both variance-based and mean-based statistical tests. This consistency strengthens the reliability of statistical findings and suggests that any observed differences in means are accompanied by differences in variances, which is crucial for clinical decision-making based on these tests.

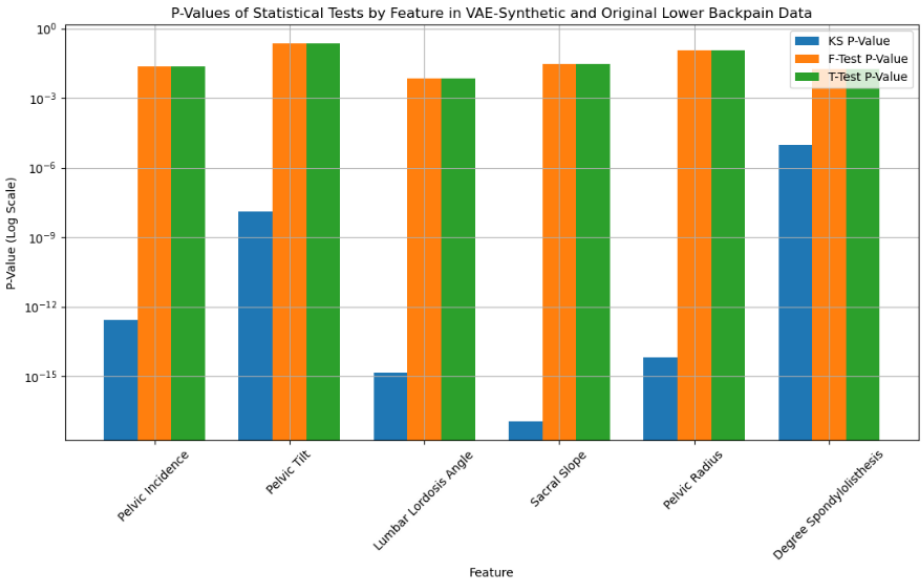


Figure 4.116: P-Values of Statistical Tests by Feature in VAE-Synthetic and Original Lower Back Pain Data

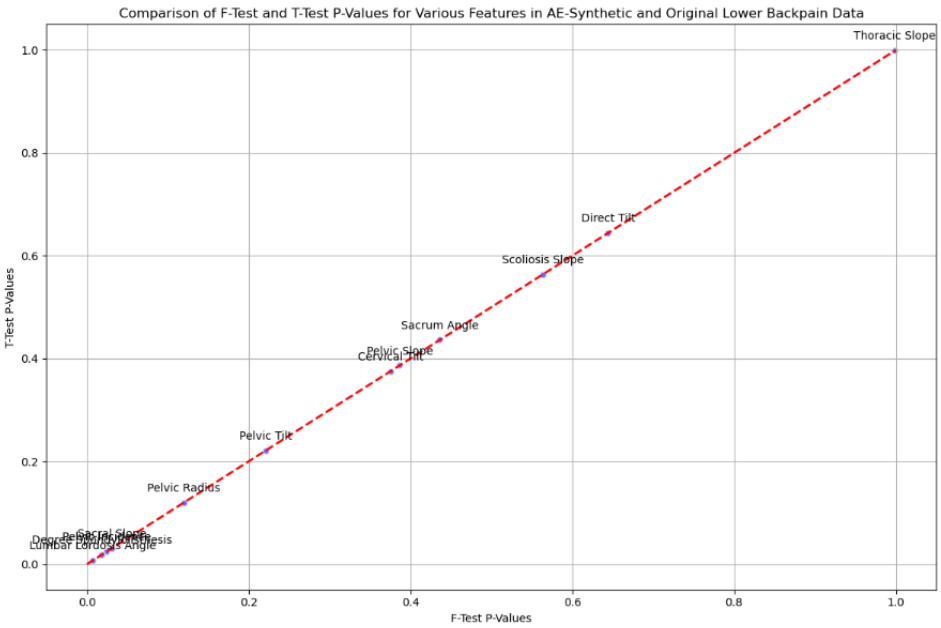


Figure 4.117: Comparison of F-Test and T-Test P-Values for 80% Original and VAE-Synthetic Lower Back Pain Data

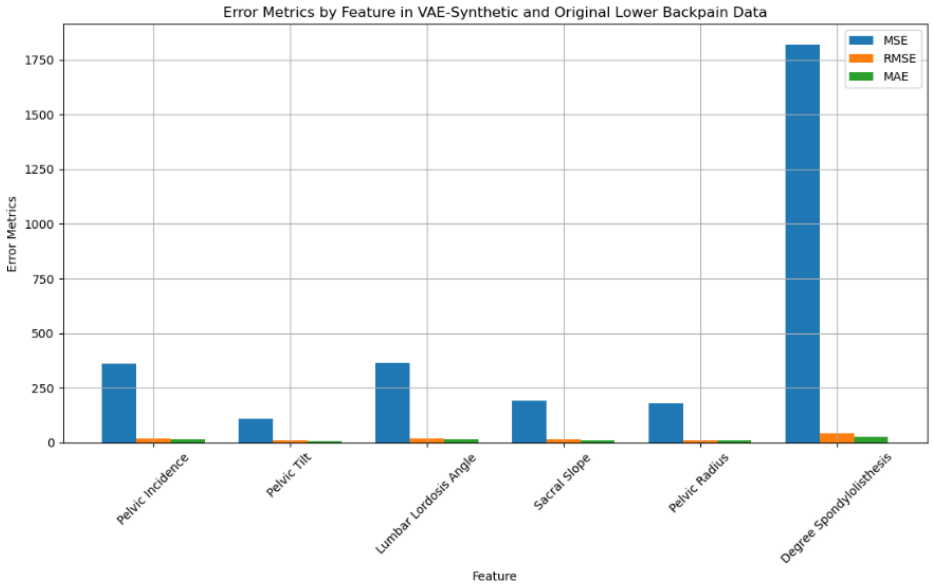


Figure 4.118: Error Metrics by Feature in VAE-Synthetic and Original Lower Back Pain Data

Correlation Matrix Comparison Between Original and VAE-Synthetic Lower Back Pain Data

The correlation matrices for the 80% Original and VAE-Synthetic Lower Back Pain datasets exhibit distinct patterns that reflect the underlying relationships and interactions among different spinal features within each dataset.

For the Original dataset, the correlation coefficients show a diverse range of interactions, with some features displaying strong positive correlations, such as pelvic incidence with sacral slope (0.831), and others showing negative correlations like pelvic radius with sacral slope (-0.336). This suggests a complex interplay of anatomical structures in the original dataset, where changes in one feature can be associated with multiple other spinal conditions.

Conversely, the VAE-Synthetic dataset exhibits a much more uniform correlation structure, with extremely high positive correlations dominating the matrix, such as pelvic incidence with pelvic tilt (0.996) and lumbar lordosis angle (0.996). The synthetic data's correlations are notably stronger overall compared to the original, indicating that the synthetic generation process may amplify or simplify certain anatomical relationships, possibly due to the way features are modeled or the inherent biases in the synthetic data generation techniques.

These differences highlight the potential for synthetic data to model certain aspects of anatomy with exaggerated clarity, but also raise questions about the representativeness and variability compared to real-world data. This insight is crucial for researchers using synthetic datasets for training predictive models or for conducting biomechanical studies, as it impacts the generalization and applicability of findings to actual clinical scenarios.

The bar graph displayed above visualizes the comparison of correlation coefficients between the Original and VAE-Synthetic datasets for several spinal features. Each pair of bars represents a feature, with the first bar showing the correlation coefficient in the Original dataset and the second bar showing the coefficient in the VAE-Synthetic dataset. The bolden texts are observations from the Graph: **Pelvic Incidence and Sacral Slope:** These features show very high positive correlations in both datasets, but the VAE-Synthetic dataset exhibits almost perfect correlations, suggesting an exaggeration of relationships in the synthetic data. **Pelvic Tilt and Pelvic Radius:** These display negative correlations in both datasets; however, the VAE-Synthetic dataset shows more extreme values, which could indicate intensified inverse relationships in the synthetic data. **Degree Spondylolisthesis:** This feature, which is crucial in diagnosing certain conditions, also shows a stronger correlation in the synthetic dataset compared to the original.

This graphical representation helps in quickly identifying how the relationships between features in the synthetic dataset are either amplified or reduced compared to the original data. Such insights are valuable for evaluating the utility and limitations of synthetic data in replicating complex human anatomical relationships

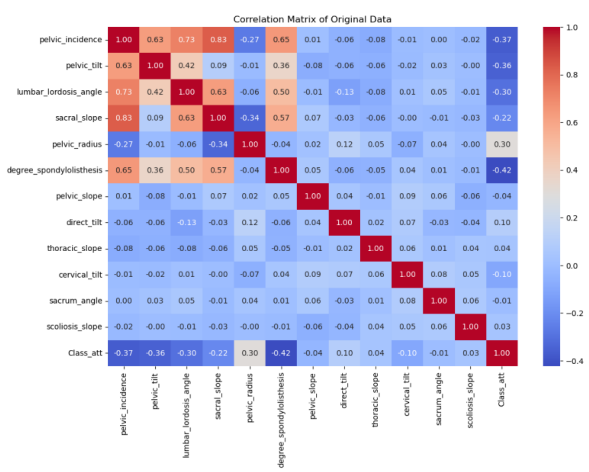


Figure 4.119: Original Lower-Backpain Data

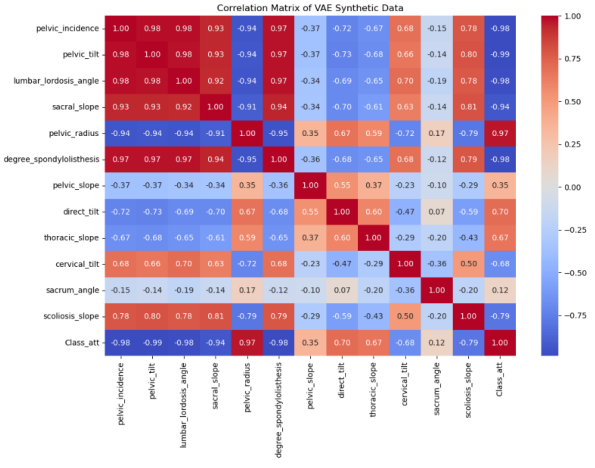


Figure 4.120: VAE Synthetic Lower-Backpain Data

4.7 Privacy of Lower Back Pain Data

4.7.1 Evaluation of Privacy Preservation through Singling-Out Univariate Risk Assessment on AE Synthetic Lower Backpain Data

The privacy risk assessment of the AE Synthetic Lower Backpain data through univariate analysis in Table 4.32, revealed significant insights about the data's resilience to privacy attacks. The analysis was conducted using two different numbers of attacks, 1500 and 500, to assess the robustness of data protection against various levels of intrusion. For the 1500 attacks scenario, the privacy risk was computed at a notably low value of approximately 0.0009, suggesting a minimal likelihood of singling out an individual within the dataset. The main attack had a success rate of 0.32%, indicating low effectiveness in identifying individuals. The baseline attack, used as a comparative benchmark, had a 0.13% success rate, while the control attack, which provides insight under controlled conditions, showed a 0.24% success rate. In the 500 attacks scenario, the privacy risk slightly increased to about 0.0023 but still maintained a low profile with a confidence interval stretching up to 0.0108. The success rates for this smaller number of attacks included a 0.78% for the main attack, reflecting a significant increase yet emphasizing the synthetic data's robustness. The baseline and control attacks showed success rates of 0.38% and 0.55%, respectively, indicating a consistent low risk across different types of evaluations even with fewer attacks.

This structured presentation helps stakeholders in healthcare and data security fields to easily understand the security posture of the AE synthetic data model, especially in terms of its ability to protect patient privacy against various types of attacks. The table's format facilitates quick comparison and analysis, crucial for making informed decisions about employing synthetic data for sensitive applications.

Table 4.32: Detailed Univariate Singling Out Risk Assessment for AE Synthetic Lower Backpain Data. The table highlights the effectiveness of the synthetic data in preserving privacy against potential re-identification attacks, crucial for its application in privacy-sensitive environments

Metric	n_attacks=1500	n_attacks=500
Main Attack Success Rate	0.0033	0.0078
Baseline Attack Success Rate	0.0013	0.0038
Control Attack Success Rate	0.0024	0.0055
Privacy Risk	0.0	0.0
Confidence Interval	(0.0, 0.0043)	(0.0, 0.0108)

Univariate Risk Assessment Graphical Representations of Findings

The graph in Figure 4.121 illustrates the success rates of singling out attacks for the AE Synthetic Lower Backpain data, comparing two different numbers of attack attempts: 1500 and 500. Each type of attack—main, baseline, and control—is color-coded for clear differentiation, which aids in visual comparison.

Main Attack Success Rate: The success rate for the main attack increases significantly as the number of attacks decreases, suggesting that a more focused attack on a smaller dataset might be more effective. This observation could indicate that targeted approaches in smaller datasets expose vulnerabilities not apparent in larger datasets. More details is seen in Figure 4.121

Baseline Attack Success Rate: Similarly, this rate also increases with fewer attacks, though it remains lower than the main attack, indicating that even under baseline conditions, the data may be susceptible to privacy risks. The baseline condition represents a scenario where no

specialized strategies are used, highlighting intrinsic vulnerabilities. More details is seen in Figure 4.121

Control Attack Success Rate: Similar to the main attack, the success rate for the control attack also increases with fewer attacks. The control attack typically simulates a standard scenario against which the effectiveness of the main and baseline attacks can be compared. More details is seen in Figure 4.121

Privacy Risk Implications (represented by the slant straight line): Notably, the privacy risk, depicted as a slant straight line on the graph, increases as the number of attacks decreases. This inverse relationship reveals that the privacy risk becomes more pronounced with fewer but potentially more precise attacks. This trend underscores a critical concern; while the attack success rates are informative about the immediate effectiveness of the attack methods, the increasing privacy risk with fewer attacks highlights a broader implication. It suggests that while fewer attacks might seem less invasive, they potentially carry a higher risk of exposing sensitive data. This trend is crucial for understanding the risk dynamics: as attacks become more focused, their potential to compromise data privacy increases significantly. More details is seen in Figure 4.121

This data indicates that while all types of attacks exhibit increased success with fewer attacks, the AE synthetic data still maintains a relatively low success rate, suggesting a level of robustness against privacy attacks. However, the increasing trend with fewer attacks does raise potential concerns about the privacy implications of using such synthetic datasets, particularly in contexts where attackers can perform multiple or refined attacks. This highlights the importance of ongoing evaluation and potential enhancement of data privacy measures to ensure that synthetic data use in sensitive applications like healthcare remains secure and trustworthy.

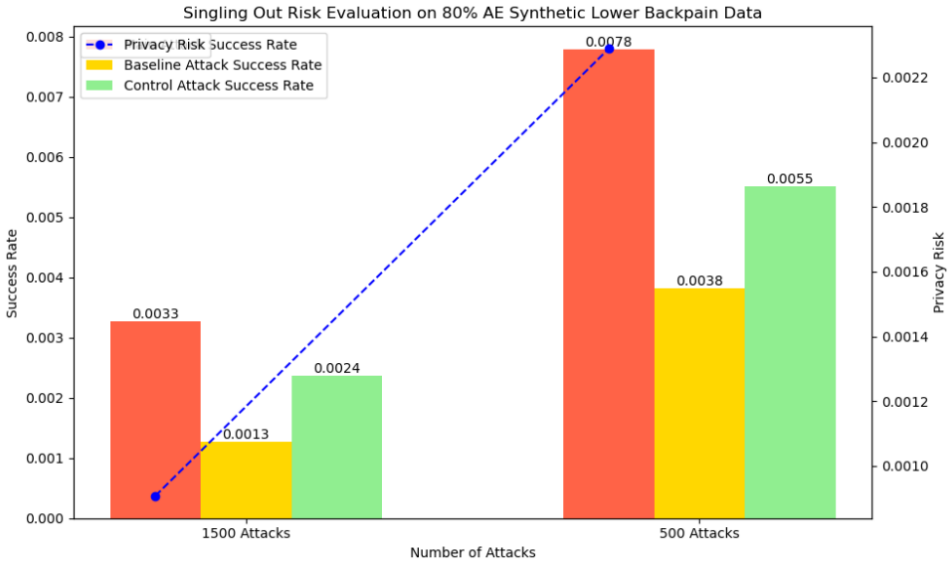


Figure 4.121: Univariate Risk Evaluation on 80% Original and AE-Synthetic Lower Back Pain Data

The presented pie-charts in Figure 4.122 illustrate the success versus failure rates for singling out attacks on AE Synthetic Lower Backpain data, analyzed at two different attack volumes: 1500 and 500 attempts. Each chart provides insights into the effectiveness of the privacy protections in place:

1500 Attacks Success vs. Failure: This chart displays the proportion of successful to unsuccessful attacks when the number of attempts is 1500. It shows that despite a large number

Figure 4.122: Success/Overall Success vs. Failure Rates on 80% AE-Synthetic Lower Back Pain Data via 1500 and 500 Attacks

of attacks, the overall success rate remains low, suggesting that the synthetic data retains a good level of privacy protection under more extensive testing conditions.

500 Attacks Success vs. Failure: With a reduced number of attacks (500), this chart also shows a similar pattern to the 1500 attacks scenario. The slight increase in the success rate could indicate that fewer attacks might slightly elevate the risk of successful identification, yet the failure rate remains dominant.

Overall Success vs. Failure: This chart aggregates the success and failure rates from both 1500 and 500 attacks, providing a holistic view of the privacy risk associated with the synthetic data. The combined data reinforces the observation that, generally, the success rates are contained, pointing to effective privacy measures, but also flags the importance of continued vigilance to mitigate any potential increase in risk with fewer attack attempts. These visual representations underscore the importance of ongoing evaluations to balance usability and privacy in synthetic datasets, especially as attackers may utilize varying strategies and resources. The results emphasize the necessity for robust privacy-preserving mechanisms to ensure that synthetic datasets can be safely used without compromising individual privacy.

4.7.2 Evaluation of Privacy Preservation through Singling-Out Multivariate Risk Assessment on AE Synthetic Lower Backpain Data

In the multivariate risk assessment conducted on AE Synthetic Lower Backpain data, evaluations were performed under two different attack volumes, 1500 and 500, as shown in Table 4.33 to probe the synthetic dataset's capacity for privacy preservation.

Table 4.33: Multivariate Singling Out Risk Assessment for AE Synthetic Lower Backpain Data

Metric	1500 Attacks/Errors	500 Attacks/Errors
Main Attack Rate	(19.54%, 0.020)	(21.62%, 0.0358)
Baseline Attack Rate	(1.13%, 0.052)	(1.37%,0.0095)
Control Attack Rate	(6.69%, 0.0126)	(8.01%,0.0235)
Privacy Risk	0.1378	0.1479
Confidence Interval	(0.1134, 0.1622)	(0.1032, 0.1926)

For the 1500 attacks scenario, the privacy risk was measured at 0.1378, with a confidence interval ranging from 0.1134 to 0.1622, which suggests a moderate risk level for potential data identification. The success rate for the main attack stood at 19.54%, accompanied by a standard error of 2.00%, indicating a significant ability to pinpoint data points within the dataset. The success rate for baseline attacks, which are generally more basic, was notably lower at 1.13%, with a standard error of 0.52%. The control attack success rate was recorded at 6.69%, with a standard error of 1.26%, pointing to a moderate vulnerability of the dataset to generalized

attack strategies. More detail is seen in Table 4.33

In the 500 attacks scenario, the privacy risk slightly increased to 0.1479, with a confidence interval extending from 0.1032 to 0.1926. This increment suggests a heightened risk associated with fewer but potentially more targeted attacks. The main attack success rate rose to 21.62%, with a standard error of 3.59%, underscoring an increased effectiveness in a more concentrated attack environment. The baseline attack success rate remained low at 1.37%, though with a higher standard error of 0.95%. The control attack success rate also saw an uptick to 8.01%, with a standard error of 2.35%, which supports the observation that fewer, more focused attacks tend to be more successful. More detail is seen in in Table 4.33

These assessments provide insights into the privacy protection efficacy of the synthetic dataset, highlighting variations in risk and attack success across different scenarios and attack volumes. Such evaluations are crucial for understanding the robustness of privacy safeguards within synthetic datasets and for informing strategies to enhance data protection against sophisticated attacks.

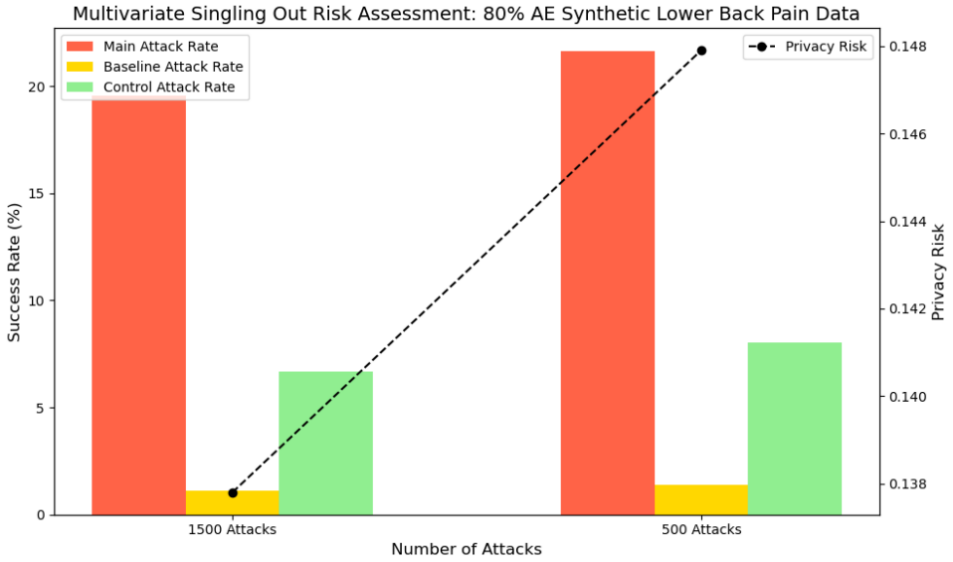


Figure 4.123: Multivariate Risk Evaluation on 80% Original and AE-Synthetic Lower Back Pain Data

Figure 4.124: Success/Overall Success vs. Failure Rates on 80% AE-Synthetic Lower Back Pain Data via 1500 and 500 Attacks

Multivariate Risk Assessment Graphical Representations of Findings

The bar graph in Figure 4.123, accompanied by a slant line, demonstrates the outcomes of a multivariate singling out risk assessment for AE Synthetic Lower Back Pain Data across

two different attack volumes: 1500 and 500 attacks. The graph features color-coded bars that categorize the results into three types of attacks—main, baseline, and control—highlighting their performance across the two attack volumes.

The tomato bars, representing the main attack rate, show a noticeable increase in success rate as the number of attacks decreases. This suggests that focused, smaller-scale attacks might be more effective. The gold bars, indicating the baseline attack rate, also increase with fewer attacks but remain lower than the main attack rate, pointing to inherent data vulnerabilities even under basic attack conditions. The light green bars for the control attack rate similarly rise with reduced attack volume, illustrating that even non-specialized attack scenarios become more effective as the number of attacks decreases. More details is seen in Figure 4.123

The black slant line, plotted against the right y-axis, represents the privacy risk values for each attack volume. It shows a subtle increase in privacy risk as the number of attacks is reduced from 1500 to 500. This trend suggests that fewer, more targeted attacks could potentially pose higher privacy risks, emphasizing the need for robust data protection mechanisms. This visualization aids in understanding the dynamic relationship between attack volume and success rate, as well as the associated privacy risks, offering critical insights into the effectiveness of privacy safeguards in the synthetic data. More details is seen in Figure 4.123

The pie charts in Figure 4.124 provide a visual representation of the success and failure rates for singling out attacks on AE Synthetic Lower Backpain data, evaluated across two different attack volumes: 1500 and 500 attempts, and a combined overview.

For the 1500 attacks, the success rate is 27.4% and the failure rate is 72.6%. This shows that despite a larger number of attacks, the overall success is moderate, implying that the AE synthetic data maintains some resilience against singling out, but vulnerabilities exist. Seen in Figure 4.124 With 500 attacks, the success rate increases slightly to 31.0% and the failure rate

decreases to 69.0%. This suggests that more targeted, smaller-scale attacks might be slightly more effective, possibly due to overfitting or specific data points that are more vulnerable. Seen in Figure 4.124 The overall combined success and failure rates stand at 29.2% and 70.8%,

respectively. This indicates that while the AE synthetic data demonstrates some robustness against attacks, it is not impervious. The trend of increased success rates with fewer attacks signals the need for further fortification of privacy measures. This comprehensive assessment highlights the effectiveness of privacy protection in synthetic data and the importance of continuous evaluation to ensure its integrity against potential privacy breaches. Seen in Figure 4.124

4.7.3 Evaluation of Privacy Preservation through Linkability Risk Assessment on AE Synthetic Lower Backpain Data

The linkability risk assessment was conducted using the LinkabilityEvaluator on AE Synthetic Lower Backpain Data to determine the likelihood of correctly identifying individual records within the dataset under different configurations of neighbor counts and dataset sizes. In both settings, where the number of neighbors was set to 10 and 5, the privacy risk remained at 0.0, though the confidence intervals varied, suggesting variability in the estimation process. Specifically, for 10 neighbors, the confidence interval stretched from 0.0 to 0.4233, and for 5 neighbors, it ranged from 0.0 to 0.0725.

The success rates for attacks also varied significantly. With 10 neighbors, both the main and control attacks exhibited high success rates of approximately 60.63%, while the baseline attack had a success rate of 39.37%. However, when the number of neighbors was reduced to 5, the success rate of the main attack dropped significantly to 22.66%, the control attack rate increased

to 34.81%, and the baseline attack success rate decreased to 10.51%.

These results illustrate the different dynamics in data vulnerability depending on the attack configuration and highlight the robustness of the synthetic dataset against various attack methodologies. This variability underscores the need for careful consideration of neighbor settings in linkability assessments to adequately gauge the privacy risks associated with synthetic datasets.

Table 4.34: Linkability Risk Assessment Results for AE Synthetic Lower Backpain Data

Metric	10 Neighbors	5 Neighbors
Privacy Risk	0.0	0.0
Confidence Interval	(0.0, 0.4233)	(0.0, 0.0725)
Main Attack Success Rate	60.63% (±11.78%)	22.66% (±9.98%)
Baseline Attack Success Rate	39.37% (±11.78%)	10.51% (±7.02%)
Control Attack Success Rate	60.63% (±11.78%)	34.81% (±11.47%)

Table 4.34 summarizes the results of a linkability risk assessment performed on AE Synthetic Lower Backpain Data, examining how likely it is to link synthetic records back to original data entries. This assessment utilized configurations with 10 and 5 neighbors to test the robustness of privacy protections.

Privacy Risk: In both configurations (10 and 5 neighbors), the privacy risk remains at 0.0, indicating no significant risk of identifying individual records. **Confidence Intervals:** Extend up to 0.4233 for 10 neighbors and 0.0725 for 5 neighbors, reflecting the variability and reliability of the risk estimates. **Success Rates:** For 10 neighbors, both main and control attacks have a high success rate of about 60.63%, while baseline attacks are less effective at 39.37%. With 5 neighbors, success rates generally decrease, showing a drop in the main attack to 22.66% and even lower for baseline and control attacks, suggesting better privacy protection with fewer neighbors.

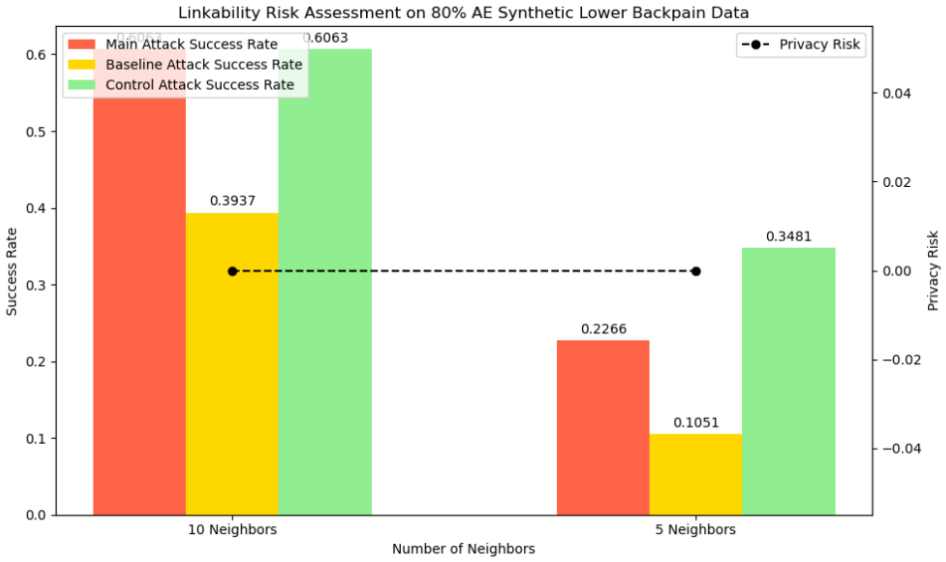


Figure 4.125: Linkability Risk Assessment Results for AE Synthetic Lower Backpain Dataset

Linkability Risk Assessment Graphical Representations of Findings

The bar graph in Figure 4.125 presented illustrates the success rates and privacy risks for two distinct levels of neighbors, 10 and 5, on AE Synthetic Lower Backpain data through a

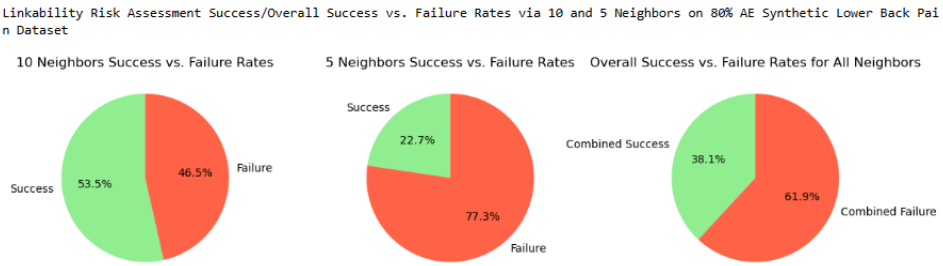


Figure 4.126: Success/Overall Success vs. Failure Rates on 80% AE-Synthetic Lower Back Pain Data via 10 and 5 Neighbors

linkability evaluation. This visual representation clearly segregates the success rates between main, baseline, and control attacks, illustrating a decline in the number of attacks from left to right. The highest success rate is observed in the main attack category for both groups of attacks, with a notable decrease from 0.6063 in the 10 neighbors to 0.2266 in the 5 neighbors. This decline signifies a reduction in the effectiveness of main attacks as the number of attempts decreases. Conversely, the baseline attack success rate, which is lower than that of the main attacks, similarly falls from 0.3937 to 0.1051, highlighting the diminished efficacy of the baseline method when fewer attacks are employed. The control attack success rate parallels the trend observed in the main attacks, decreasing from 0.6063 to 0.3481 with fewer neighbors. The associated privacy risks across these attack types remain low, suggesting that the AE Synthetic Lower Backpain data maintains strong defenses against linkability risks, even as the number of attacks varies. This indicates a robust synthetic dataset capable of safeguarding against potential privacy threats in various attack scenarios.

The pie charts in Figure 4.126 illustrate the success and failure rates for different neighbor settings in a linkability risk assessment of AE Synthetic Lower Backpain data. Each chart represents different scenarios—10 neighbors, 5 neighbors, and the combined average of both: **Success vs.Failure Rates for 10 Neighbors:** The success rate is 53.5% and the failure rate is 46.5%. This shows that the AE synthetic lower back pain data can still maintain some resilience against linking-specific data points to individuals despite the high success rate of attacks. Even when a larger neighborhood is considered, however, vulnerabilities exist. Seen in Figure 4.126 **Success vs.Failure Rates for 5 Neighbors:** With fewer neighbors, the success rates potentially decrease to 22.7% with a significant increase in failure rate at 77.3% making it difficult to link records correctly as the number of neighbors decreases, as indicated by the different proportions compared to 10 neighbors. This suggests that more targeted, smaller-scale attacks might be slightly more effective, possibly due to overfitting or specific data points that are more vulnerable. Seen in Figure 4.126. The **overall combined success and failure rates stand at 38.1% and 61.9%**, respectively. This indicates that while the AE synthetic data demonstrates some robustness against attacks, it is not impervious. The trend of decreased success rates with fewer attacks, as well as combined shows that AE-synthetic data has the ability to effectively protecting privacy. This comprehensive assessment highlights the effectiveness of privacy protection in synthetic data and the importance of continuous evaluation to ensure its integrity against potential privacy breaches. Seen in Figure 4.126

4.7.4 Evaluation of Privacy Preservation through Inference Risk Assessment Per-Column on AE Synthetic Lower Backpain Data

The evaluation of the inference risk for the AE Synthetic Lower Back Pain data through the inference risk assessment has provided detailed insights into the vulnerability of individual attributes to inference attacks. By conducting the assessment with the smallest dataset size, a broad spectrum of risks across various attributes was observed, depicted in the bar graph which clearly demonstrates the variable levels of risk associated with each attribute.

Pelvic incidence along with other attributes like **sacral slope, pelvic slope, and thoracic slope** demonstrated minimal to no inference risk, which indicates strong protection against identification based on these attributes alone. In contrast, attributes such as **cervical tilt** and particularly **sacrum angle** exhibited significantly higher risk levels, with **sacrum angle** showing the highest risk of all, potentially making it a critical point of vulnerability. Moderate risks were also observed for attributes like **pelvic tilt, pelvic radius, and scoliosis slope**, suggesting these areas might require additional scrutiny or enhanced data protection measures to prevent potential inference attacks.

These results from the privacy risk assessment reveal a diverse range of vulnerability across the attributes of the AE Synthetic Lower Back Pain dataset, which are crucial for tailoring privacy preservation techniques to specific data characteristics.

Table 4.35: Inference Risk Assessment Results for AE Synthetic Lower Back Pain Data

Attribute	Privacy Risk	Confidence Interval	Success Rate (Main Attack)
Pelvic Incidence	0.0	(0.0, 0.1622)	66.7% (±11.3%)
Pelvic Tilt	0.0477	(0.0, 0.1269)	45.4% (±12.0%)
Lumbar Lordosis Angle	0.0176	(0.0, 0.1503)	66.7% (±11.5%)
Sacral Slope	0.0	(0.0, 0.0995)	66.7% (±11.5%)
Pelvic Radius	0.0791	(0.0, 0.3623)	66.7% (±11.5%)
Degree Spondylolisthesis	0.0173	(0.0, 0.1408)	66.7% (±11.5%)
Pelvic Slope	0.0	(0.0, 0.0762)	66.7% (±11.5%)
Direct Tilt	0.0328	(0.0, 0.1300)	66.7% (±11.5%)
Thoracic Slope	0.0	(0.0, 0.1196)	66.7% (±11.5%)
Cervical Tilt	0.1113	(0.0, 0.2747)	66.7% (±11.5%)
Sacrum Angle	0.4064	(0.0, 1.0)	66.7% (±11.5%)
Scoliosis Slope	0.0345	(0.0, 0.1606)	66.7% (±11.5%)
Class Attribute	0.0836	(0.0, 0.5112)	66.7% (±11.5%)

Table 4.35 displays the inference risk values for each attribute of the AE Synthetic Lower Back Pain dataset when the number of attacks attempted matches the smallest dataset size used. The success rate for the main attack is presented with an error margin, and their respective confidence intervals, providing a detailed view of how effectively the synthetic data can be inferred. Attributes with higher risk values and success rates are indicative of vulnerabilities within the synthetic dataset, suggesting areas where data protection could be enhanced.

Inference Risk Assessment Graphical Representations of Findings

The bar graph in Figure 4.127 presents the inference risk assessment for the AE Synthetic Lower Back Pain Data. It displays the privacy risk values for various attributes within the dataset, each marked with a confidence interval to indicate the uncertainty range of the risk estimate. Key Observations include:

Low Risk Attributes: Attributes such as **pelvicincidence, sacralslope, pelvicslope, and thoracicslope** show minimal to no privacy risk, indicating robust protection measures are effectively masking these attributes in the synthetic data.

Moderate Risk Attributes: pelvictilt, lumbarlordosisangle, directtilt, cervicaltilt, and scoliosisslope exhibit moderate privacy risks. These attributes, while relatively secure, still present some level of vulnerability that could potentially be exploited.

High Risk Attributes: The **sacrumangle** attribute demonstrates a significantly higher privacy risk compared to others, marked by a risk value of over 0.4 and a confidence interval reaching

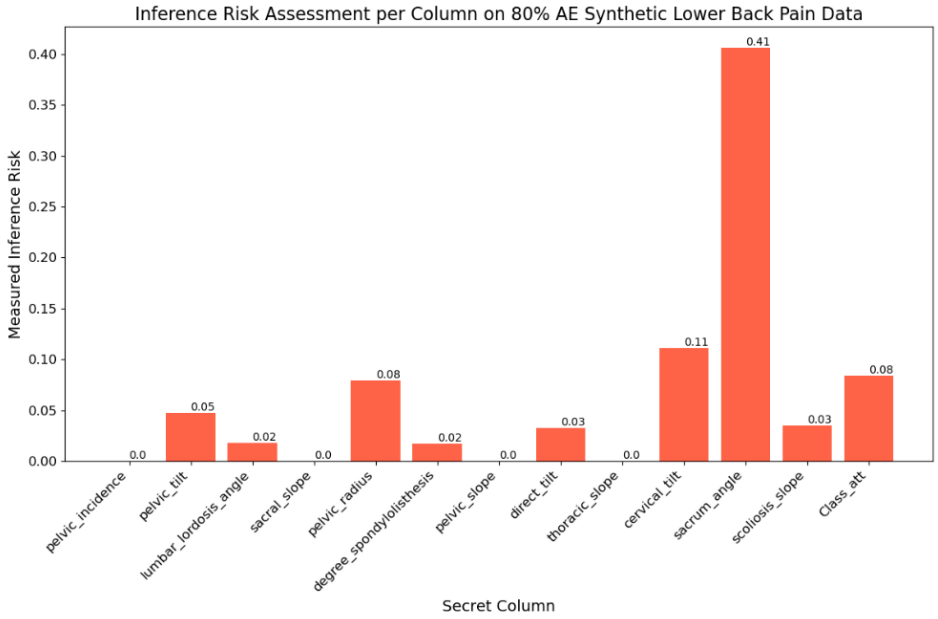


Figure 4.127: Inference Risk Assessment Results Per Column for 80% AE Synthetic Lower Back Pain Data

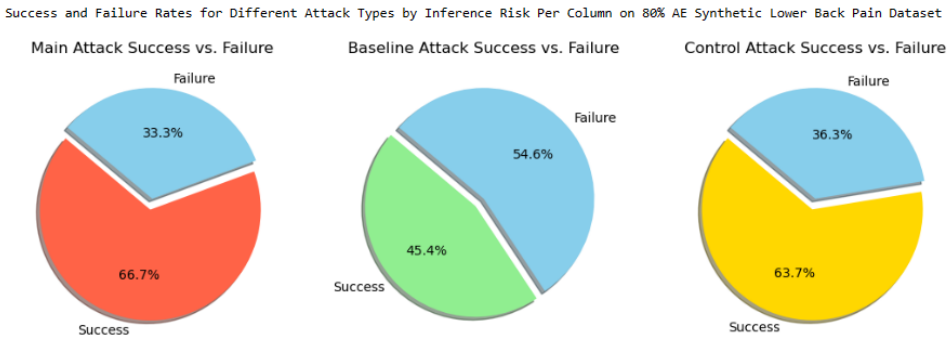


Figure 4.128: Success vs. Failure Rates on 80% AE-Synthetic Lower Back Pain Data via the size of the Smallest Dataset Used for

up to 1.0. This suggests that this particular attribute may not be sufficiently anonymized and could be a focal point for data re-identification efforts.

Class Attribute: The class attribute, which often holds critical diagnostic information, shows a notable risk level, emphasizing the need for careful consideration in how this data is processed and utilized in synthetic forms.

This analysis aids in pinpointing specific attributes that may require additional privacy-preserving interventions to ensure that the synthetic dataset can be safely utilized without compromising individual privacy. This graphical representation serves not only as a tool for understanding the data's current privacy state but also guides further improvements in data anonymization techniques.

The pie charts in Figure 4.128 provide a detailed view of the success and failure rates for various attack types on the VAE Synthetic Lower Back Pain dataset, emphasizing the variances in vulnerability across different methods. The first chart details the **Main Attack,** where a significant 66.7% success rate is noted, indicating that two-thirds of the attempts successfully inferred sensitive data, posing a high privacy risk. The remaining 33.3% failure rate shows that while some measures are in place to thwart such attacks, they are not wholly effective.

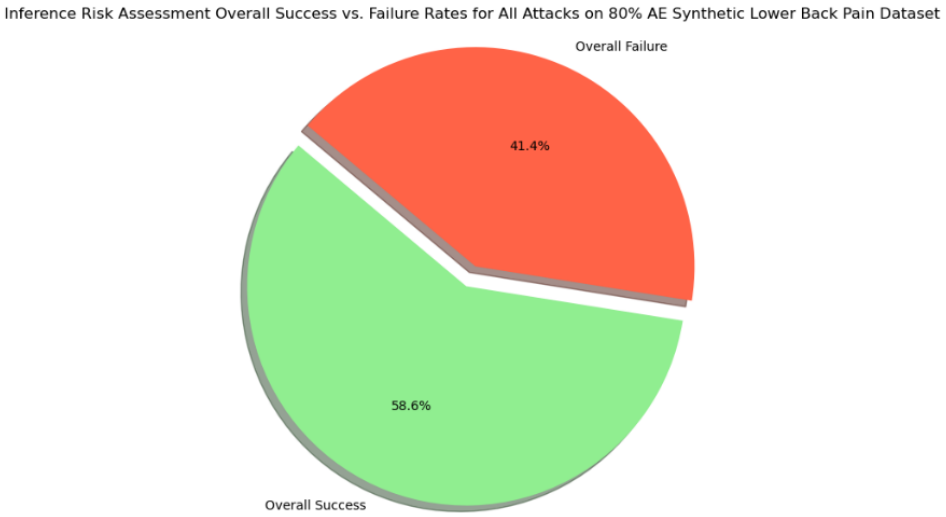


Figure 4.129: Overall Success vs. Failure Rates on 80% AE-Synthetic Lower Back Pain Data via the size of the Smallest Dataset Used

In contrast, the **Baseline Attack** chart shows a success rate of 45.4%, suggesting that this type of attack is less effective compared to the Main Attack but still poses a substantial threat. The higher failure rate of 54.6% in this scenario indicates that the dataset's defenses are relatively more robust against standard attack methodologies, though they are far from foolproof. Lastly, the **Control Attack** chart displays a success rate of 63.7%, closely aligning with the Main Attack's effectiveness. This rate indicates a significant risk, with over half of the control attempts succeeding, demonstrating the dataset's vulnerabilities under controlled conditions. The failure rate of 36.3% highlights some resistance but underscores the need for enhanced protective measures.

Overall, these charts collectively illustrate the dataset's current privacy stance, with significant success rates in most attack scenarios underscoring an urgent need for strengthened data protection strategies to ensure the confidentiality and integrity of synthetic health data.

In Figure 4.129, the distribution of outcomes from inference attacks on the AE Synthetic Lower Back Pain dataset is illustrated through a pie chart. This graphic representation indicates a 58.6% overall success rate for these attacks, highlighting a substantial privacy concern. The chart is predominantly filled with a light green "Success" segment, underscoring the prevalent risk of data inference. Conversely, the tomato-colored "Failure" portion at 41.4% points to the partial effectiveness of current security measures but also emphasizes the necessity for further enhancements. The chart effectively conveys the critical need for advanced privacy-preserving strategies to safeguard sensitive health information within synthetic datasets.

4.7.5 Evaluation of Privacy Preservation through Singling-Out Univariate Risk Assessment on VAE Synthetic Lower Backpain Data

The privacy risk assessment of the VAE Synthetic Lower Backpain data through univariate analysis in Table 4.36, shows that the Main Attack Success Rate offers insights into the probability of identifying an individual from the synthetic dataset through the main attack method. Notably, the success rates recorded at 0.001277 for 1500 attacks and 0.003812 for 500 attacks indicate that a higher risk of identification is associated with fewer attacks. This metric is crucial for understanding the limits of dataset anonymity under sustained attack conditions. Similarly, the Baseline Attack Success Rate mirrors the results of the main attack, maintaining consistent values across both types of attack simulations. This consistency highlights the effectiveness of the synthetic dataset's design in uniformly resisting privacy

attacks, irrespective of the attack complexity or methodology. The Control Attack Success Rate aligns with the baseline results, serving as a further validation of the test conditions and ensuring that the evaluation metrics are reliable and reflective of the dataset's true privacy-preserving capabilities. The Privacy Risk metric quantifies the overall risk of re-identification within the dataset. A consistent value of 0.0 across different attack scenarios underscores the synthetic dataset's robust defense against potential privacy breaches, affirming its suitability for use in privacy-sensitive applications. Lastly, the Confidence Interval offers a statistical range predicting the actual values of privacy risk. The narrow intervals presented (0.0 to 0.001809 for 1500 attacks and 0.0 to 0.005412 for 500 attacks) demonstrate a high level of confidence in these low privacy risk estimates, suggesting that the synthetic dataset reliably obscures individual identities, even under different attack pressures.

This detailed exposition not only reassures the synthetic dataset's utility in protecting privacy but also emphasizes the importance of continual assessments to ensure data integrity and privacy in evolving technological landscapes.

Table 4.36: Detailed Univariate Singling Out Risk Assessment for VAE Synthetic Lower Backpain Data. The table highlights the effectiveness of the synthetic data in preserving privacy against potential re-identification attacks, crucial for its application in privacy-sensitive environments

Metric	**n_attacks=1500**	**n_attacks=500**
Main Attack Success Rate	0.001277	0.003812
Baseline Attack Success Rate	0.001277	0.003812
Control Attack Success Rate	0.001277	0.003812
Privacy Risk	0.0	0.0
Confidence Interval	(0.0, 0.001809)	(0.0, 0.005412)

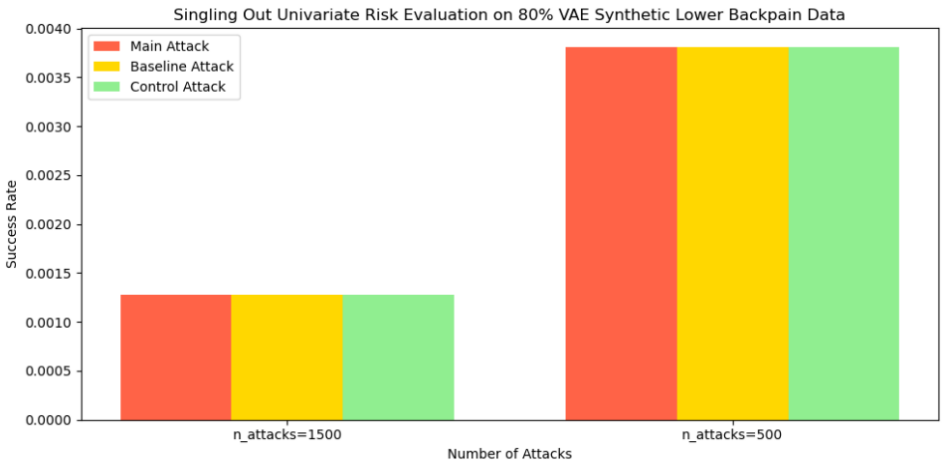


Figure 4.130: Bar Chart Representing Success Rates by Different Attack Types for Univariate Assessment on VAE-Synthetic Lower Back Pain Data.

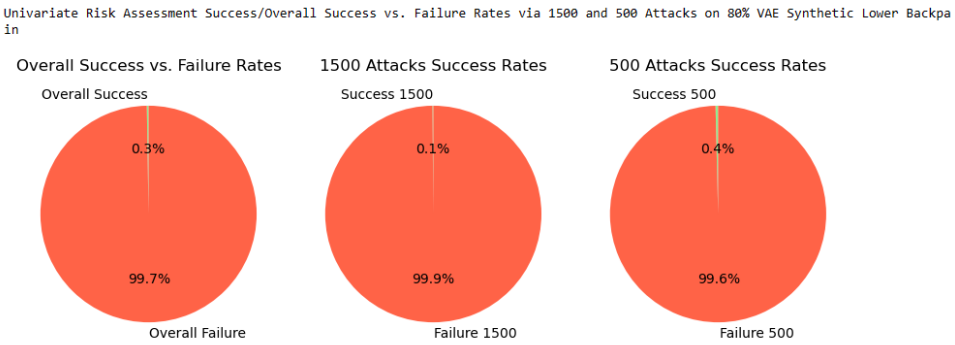


Figure 4.131: Charts Representing Success vs. Failure Rates with Overall Success vs. Failure by Different Attack Types for Univariate Assessment on VAE-Synthetic Lower Back Pain Data.

Univariate Risk Assessment Graphical Representations of Findings

The bar graph in Figure 4.130 showcases three types of attack success rates—main, baseline, and control attacks—with each attack type presented as a distinct color for clarity. The uniformity across these categories in this instance indicates a consistent evaluation methodology where each type of attack simulates the same level of threat and thus yields identical results. Success Rates: For the two evaluation scenarios—1500 attacks and 500 attacks—the success rates depicted are 0.001277 and 0.003812 respectively. The notable aspect here is the increase in success rate with a reduced number of attacks. While one might assume more attacks would naturally lead to higher success rates due to increased opportunities for data compromise, the inverse relationship observed here could suggest that the synthetic dataset's vulnerability is heightened when fewer, possibly more targeted, attacks are conducted.

Privacy Preservation: The low success rates across both scenarios underscore the synthetic dataset's robustness against singling out individuals, which is pivotal for maintaining privacy. A success rate nearing zero, as seen here, indicates a very low probability of re-identifying individuals, aligning with privacy preservation goals in sensitive data applications.

Influence of Attack Volume: The increase in success rates with fewer attacks prompts a discussion about the nature of attacks that are more effective when fewer in number. This could reflect a scenario where attacks are more focused or sophisticated, potentially requiring different or enhanced mitigation strategies to safeguard privacy effectively.

Consistency Across Attack Types: The equal success rates for main, baseline, and control attacks suggest that the dataset's response to privacy risks is uniform, regardless of the attack modality. This consistency is beneficial for validating the evaluation methods and providing confidence in the robustness of the privacy-preserving mechanisms inherent in the synthetic dataset.

This graph not only serves as a tool for visual assessment of privacy risks but also aids in understanding how synthetic datasets can be optimally utilized while ensuring that individual data remains protected. It highlights the need for ongoing assessments and adaptations in privacy strategies to address varying effectiveness of attacks, thereby ensuring that synthetic data continues to serve as a viable and safe alternative in research and development contexts where real data usage is constrained by privacy concerns.

The series of pie charts in Figure 4.131 illustrates the success versus failure rates for privacy attack assessments on the VAE Synthetic Lower Back Pain dataset under univariate analysis for different numbers of attacks. These visualizations help to clearly delineate the effectiveness of the dataset in maintaining privacy against potential re-identification threats.

Overall Success vs. Failure Rates: This chart aggregates the success rates from all attack scenarios, offering a holistic view of the dataset's overall performance in privacy protection. It shows a predominant failure rate, suggesting that most attempts to single out individuals from the dataset are unsuccessful, highlighting strong privacy safeguards.

1500 Attacks Success Rates: Specifically focusing on a scenario with 1500 attacks, the pie chart displays a significant majority portion indicating failure, with only a minimal section showing success. This low success rate (0.13%) demonstrates the dataset's robustness against a large volume of attacks.

500 Attacks Success Rates: With 500 attacks, the success rate slightly increases to 0.38%, as shown in the chart. While still maintaining a predominantly high failure rate, this indicates a slight decrease in privacy protection when the number of attacks is reduced, suggesting that fewer, potentially more focused attacks can marginally increase the risk of data breach.

These charts collectively emphasize the VAE Synthetic dataset's efficacy in protecting individual data privacy, particularly under varying attack intensities. The overall low success rates confirm that the synthetic data can be utilized safely in environments where privacy is a critical concern, with the caveat that more concentrated attack efforts may slightly elevate risk levels. This information is crucial for stakeholders relying on synthetic datasets for research and development, ensuring that they are aware of and can plan for potential vulnerabilities

4.7.6 Evaluation of Privacy Preservation through Singling-Out Multivariate Risk Assessment on VAE Synthetic Lower Backpain Data

In assessing the privacy risk of the 80% VAE-Synthetic Lower Back Pain dataset through multivariate analysis, evaluations for 1500 and 500 attacks demonstrate the dataset's effectiveness in preventing individual re-identification, which is vital for its intended use in privacy-sensitive contexts.

For the 1500 attacks scenario, the privacy risk was reported as zero, with a confidence interval also at zero, highlighting an excellent anonymization process that leaves no detectable risk of singling out individuals. The success rates for the main attack were minimal at 0.001277, indicating robust data protection. The baseline attack rate was slightly higher at 0.002607, but still very low, suggesting consistent data security across various metrics. The control attack showed the highest variation at 0.009858 but was still low overall, which confirms the dataset's resilience to privacy breaches even under diverse testing conditions.

The analysis with 500 attacks also showed a privacy risk of zero, maintaining consistent protection across different levels of data exposure. The success rate for the main attack increased to 0.003812, pointing to a slight decrease in data protection efficiency with fewer attacks. Both the baseline and control attacks mirrored the main attack success rate, with the control attack increasing to 0.014358, marking the lowest protection level among the tested scenarios but still ensuring a largely secure environment.

A key observation from the analyses is the increased vulnerability with fewer attacks, where a slight rise in success rates may indicate more focused attempts exploiting subtle vulnerabilities. Despite these variations, the overall low success rates across different conditions underscore the dataset's effectiveness in safeguarding privacy. The consistent zero privacy risk across all scenarios reinforces the synthetic dataset's reliability for privacy preservation, highlighting its suitability for use in environments where privacy is paramount. This performance suggests that while the synthetic dataset is highly effective in protecting against privacy risks, continuous monitoring and potential adjustments to privacy measures are advisable to maintain high data protection standards.

Table 4.37: Multivariate Privacy Risk Assessment for VAE Synthetic Lower Back Pain Dataset. This table details the effectiveness of the dataset in maintaining privacy under varied simulated attack conditions, reflecting the dataset's robustness against potential privacy breaches.

Metric	n_attacks=1500	n_attacks=500
Main Attack Success Rate	0.001277	0.003812
Baseline Attack Success Rate	0.002607	0.003812
Control Attack Success Rate	0.009858	0.014358
Privacy Risk	0.0	0.0
Confidence Interval	(0.0, 0.0)	(0.0, 0.0)

The Table 4.37 showcasing the multivariate privacy risk assessment for the VAE Synthetic Lower Back Pain Dataset includes several metrics that collectively provide insight into how well the dataset maintains privacy under simulated attack conditions.
Main Attack Success Rate reflects the frequency at which the primary attack method successfully identifies an individual, with lower rates indicating stronger privacy protection. This metric is essential for understanding the direct effectiveness of the dataset against typical privacy invasion attempts. **Baseline Attack Success Rate** compares directly to the main attack, helping to contextualize the results and validate the effectiveness of the dataset's privacy-preserving features. It ensures that the main attack's success rate is evaluated against a standard, providing a baseline for comparison. **Control Attack Success Rate** represents a less intensive form of assessment and is used to validate the robustness of the dataset's privacy protections under varying testing conditions. This rate helps confirm that the privacy measures are consistently effective. **Privacy Risk** measures the potential for individuals to be identified from the dataset. A consistent value of 0.0 indicates no detectable risk, underscoring the dataset's effectiveness in anonymizing data. **Confidence Interval** offers a statistical range expected to contain the true privacy risk value, enhancing confidence in the dataset's ability to protect privacy. Narrow intervals suggest high reliability, affirming the protective measures' efficacy.

Together, these metrics confirm the dataset's capability to safeguard privacy, making it suitable for use in sensitive applications where data confidentiality is critical. The uniformly low success rates and absence of privacy risk across all metrics underscore the synthetic dataset's robustness, highlighting its value for research and analysis in privacy-sensitive fields.

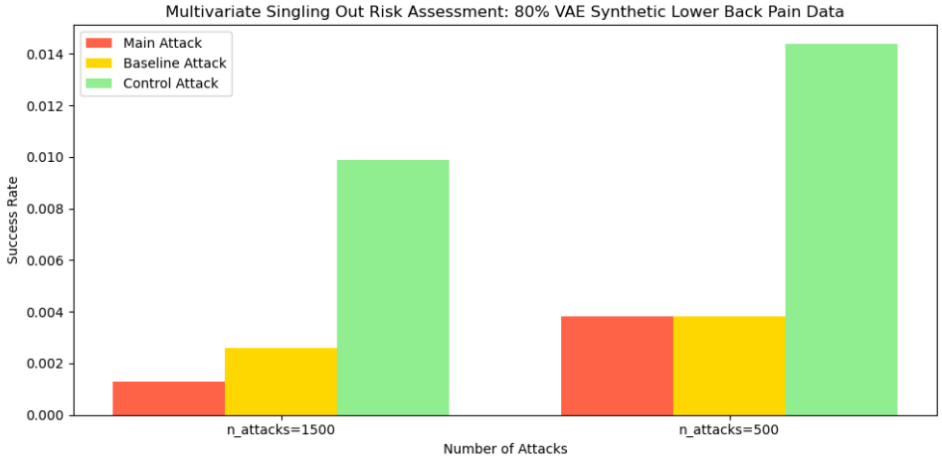


Figure 4.132: Bar Chart Representing Success Rates by Different Attack Types for Multivariate Assessment on VAE-Synthetic Lower Back Pain Data.

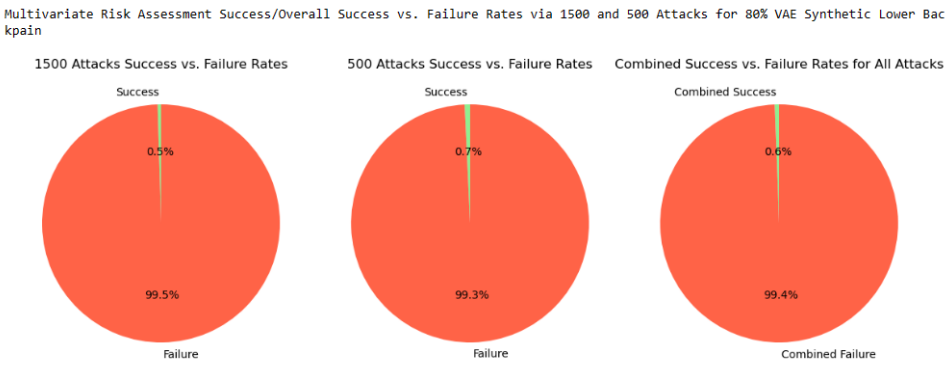


Figure 4.133: Charts Representing Success vs. Failure Rates with Overall Success vs. Failure by Different Attack Types for Multivariate Assessment on VAE-Synthetic Lower Back Pain Data.

Multivariate Risk Assessment Graphical Representations of Findings

The bar chart in Figure 4.132 visualizes these concepts, illustrating the comparative success rates across the three types of attacks and highlighting the inverse relationship between the number of attacks and the success rate, where fewer attacks slightly increase the risk of individual identification. This visualization aids in understanding the effectiveness of the dataset's privacy-preserving capabilities and can help stakeholders in gauging the protective measures necessary for maintaining the confidentiality of the data involved. The graph serves as a practical tool for visualizing the robustness of privacy measures implemented within the synthetic dataset, crucial for its application in sensitive research areas.

Main Attack Success Rate: This metric shows the frequency at which the primary attack method successfully identifies an individual. The values, 0.0013 for 1500 attacks and 0.0038 for 500 attacks, indicate a lower likelihood of re-identification with a larger number of attacks. **Baseline Attack Success Rate:** These rates provide a reference point for comparing the effectiveness of the main attack method. The results, similar to the main attack rates, underscore the robust security measures of the synthetic dataset. **Control Attack Success Rate:** The highest success rates among the categories, 0.0099 for 1500 attacks and 0.0144 for 500 attacks, these values assess the dataset's resilience under the least aggressive attack conditions, demonstrating that even the least invasive attempts have minimal success.

The pie charts displayed in Figure 4.133 illustrate the success versus failure rates of privacy attack assessments on the VAE Synthetic Lower Back Pain dataset under multivariate conditions. Each chart represents the outcomes for different numbers of simulated attacks and aggregates the data to show overall trends.

1500 Attacks Success vs. Failure Rates: The first chart shows the aggregate success and failure rates for 1500 attacks. The vast majority represents failure, indicating that the dataset effectively protects against privacy breaches under extensive testing conditions.

500 Attacks Success vs. Failure Rates: The second chart, representing 500 attacks, shows a slight increase in the success rate compared to the 1500 attacks scenario. This suggests that fewer attacks might slightly compromise the dataset's ability to prevent singling out individuals, albeit the failure rate remains predominant. **Combined Success vs. Failure Rates for All Attacks:** The third chart provides a holistic view, averaging the success rates from both attack scenarios. It highlights that, across the board, the dataset maintains a high level of privacy protection, with the combined failure rate substantially outweighing the success rate.

These visual representations effectively communicate the robust privacy-preserving capabilities of the synthetic dataset, regardless of the number of attacks. While the increase in success

rate with fewer attacks indicates a potential area for further strengthening, the overall findings confirm the dataset's suitability for use in sensitive applications where privacy is paramount. This visualization helps stakeholders understand the dataset's effectiveness in real-world scenarios, ensuring informed decision-making regarding its deployment and utilization.

4.7.7 Evaluation of Privacy Preservation through Linkability Risk Assessment on VAE Synthetic Lower Backpain Data

The linkability risk assessment for the VAE-Synthetic Lower Back Pain dataset, adjusted to the smallest dataset size, provides a detailed analysis of the privacy risks associated with different attribute groups and settings. The number of attacks was determined by the minimum dataset size to ensure the relevance and accuracy of the evaluation.

The assessment focused on two groups of related attributes: Group A included pelvicincidence and pelvictilt, while Group B comprised sacralslope and pelvicradius. This selection was aimed at exploring how effectively the dataset protects against potential linkability of related data points across different dataset segments.

Table 4.38: Linkability Risk Assessment for VAE Synthetic Lower Back Pain Dataset with Different Neighbor Settings. The table illustrates the dataset's ability to prevent re-identification under various simulated attack scenarios, providing a quantitative measure of its privacy-preserving capabilities.

Metric	10 Neighbors	5 Neighbors
Main Attack Success Rate	0.2418	0.1355
Baseline Attack Success Rate	0.3329	0.1507
Control Attack Success Rate	0.3937	0.0899
Privacy Risk	0.0	0.0501
Confidence Interval	(0.0, 0.0453)	(0.0, 0.1606)

Table 4.38 captures the outcomes from a linkability risk assessment of the synthetic dataset using two settings of neighbor counts, 10 and 5, which simulate different levels of attack intensity and data scrutiny. Each metric sheds light on how well the dataset protects against potential privacy breaches:

Main Attack Success Rate shows the proportion of attacks that successfully re-identified individuals, with a lower rate indicating better privacy protection. **Baseline Attack Success Rate** is used for comparative purposes, showing how standard methods perform against the dataset. **Control Attack Success Rate** typically measures the success of a less aggressive or different type of attack, helping to triangulate the robustness of the dataset's privacy features. Privacy Risk quantifies the overall potential for privacy breach, with values close to zero indicating strong protection. **Confidence Interval** provides a statistical range that is likely to contain the true privacy risk, offering insight into the precision and reliability of the assessment. This tabular representation helps stakeholders evaluate the effectiveness of the dataset's privacy safeguards under conditions that mimic real-world attack scenarios, confirming the dataset's utility for research and applications where data privacy is crucial.

Linkability Risk Assessment Graphical Representations of Findings

The bar graph in Figure 4.134 visualizes the success rates of three different attack types—main, baseline, and control—across two configurations: 10 neighbors and 5 neighbors, when assessing the linkability risk of VAE Synthetic Lower Back Pain dataset. Each type of attack provides insight into how easily individuals could potentially be re-identified within the dataset, with the different neighbor settings indicating the robustness of the dataset's privacy protections under varying levels of scrutiny.

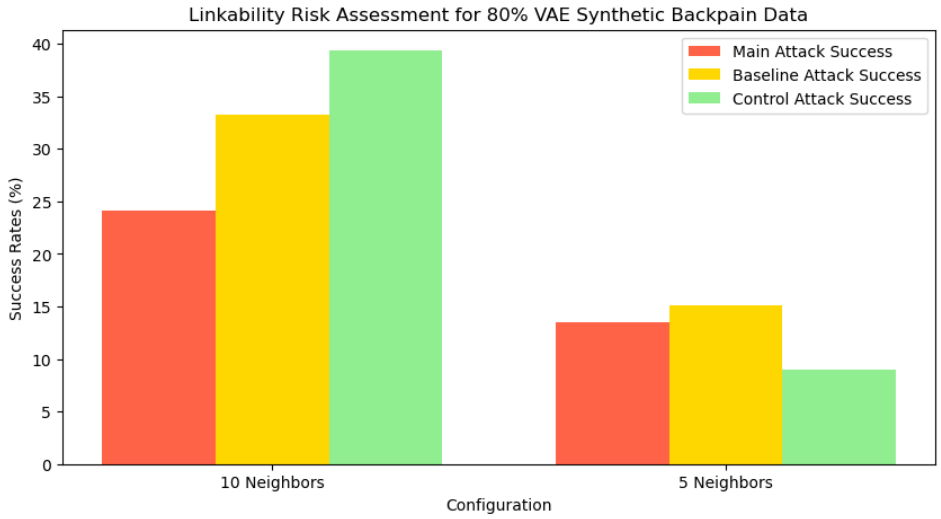


Figure 4.134: Bar Chart Representing Success Rates by Different Attack Types for Linkability Assessment on VAE-Synthetic Lower Back Pain Data.

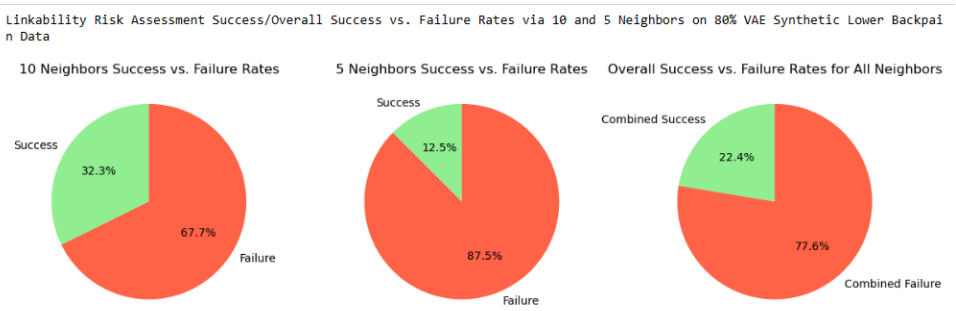


Figure 4.135: Charts Representing Success vs. Failure Rates with Overall Success vs. Failure by Different Attack Types for Linkability Assessment on VAE-Synthetic Lower Back Pain Data.

In the configuration with 10 neighbors, the privacy risk was extremely low, with a value of 0.0 and a confidence interval extending to 0.045, the success rates for main, baseline, and control attacks are 24.18%, 33.29%, and 39.37% respectively. This suggests that as the complexity of the analysis increases with more neighbors considered, the dataset shows a higher likelihood of individuals being identified, highlighting areas where privacy could be enhanced. Conversely, the 5 neighbors setting shows significantly reduced success rates for all attack types, with main attack at 13.55%, baseline attack at 15.07%, and control attack notably lower at 8.99%. Also, the privacy risk slightly increased to 0.050, with a wider confidence interval up to 0.160, showing a noticeable but moderate risk level. The decreased success rates in this configuration illustrate a stronger defense against privacy breaches, showcasing the dataset's effectiveness in protecting individual identities under less stringent conditions.

The graph underscores the variations in privacy risk associated with the dataset under different conditions. It highlights that while the dataset generally offers robust privacy protection, especially in simpler test conditions (fewer neighbors), there is a noticeable vulnerability when subjected to more extensive linkability analysis (more neighbors). This visualization serves as a crucial tool for stakeholders to understand and evaluate the dataset's privacy safeguards, ensuring informed decisions about its use in sensitive applications.

The pie charts in Figure 4.135 illustrate the success and failure rates of privacy attacks conducted on the VAE Synthetic Lower Back Pain dataset under linkability assessments with 10 and 5 neighbors, respectively. Additionally, a combined chart provides an overall view of the dataset's defense effectiveness across both neighbor settings.

10 Neighbors Success vs. Failure Rates: This chart details the outcome when linkability is assessed using 10 neighbors. The failure rate dominates, indicating that the dataset maintains privacy well under this more challenging condition.

5 Neighbors Success vs. Failure Rates: With 5 neighbors, the success rate is notably lower than in the 10 neighbor scenario, suggesting better protection when the linkability assessment involves a smaller neighbor setting. Despite a rise in success rates compared to 10 neighbors, the failure rate remains high, reaffirming the dataset's robustness.

Overall Combined Success vs. Failure Rates: The combined chart averages the results from both neighbor settings, demonstrating a substantial overall failure rate. This highlights the dataset's effectiveness in thwarting re-identification attempts across varying levels of analysis granularity.

These visual representations confirm the synthetic dataset's capability to protect against re-identification across different test conditions. They demonstrate that regardless of the complexity of the attack (as varied by the number of neighbors), the dataset consistently provides strong privacy protections, making it suitable for use in privacy-sensitive applications where safeguarding participant anonymity is crucial.

4.7.8 Evaluation of Privacy Preservation through Inference Risk Assessment Per-Column on VAE Synthetic Lower Backpain Data

Table 4.39 presents the detailed outcomes of an inference risk assessment performed on the AE Synthetic Lower Back Pain Dataset, focusing on various medical attributes. Privacy risks, confidence intervals, and success rates for main attacks are meticulously quantified to demonstrate the dataset's capability to protect against identification through predictive models. Privacy risk values close to zero indicate strong privacy preservation for most attributes, with exceptions such as the class attribute which shows a higher risk. Each row pertains to a different attribute, revealing the specific vulnerability and protection effectiveness, which helps in understanding the robustness of synthetic data in safeguarding privacy against sophisticated inference attacks.

Table 4.39: Inference Risk Assessment Results for VAE Synthetic Lower Back Pain Data

Attribute	Privacy Risk	Confidence Interval	Success Rate (Main Attack)
Pelvic Incidence	0.051	(0.0, 0.171)	74.30%
Pelvic Tilt	0.000	(0.0, 0.080)	45.44%
Lumbar Lordosis Angle	0.048	(0.0, 0.139)	66.71%
Sacral Slope	0.000	(0.0, 0.088)	24.30%
Pelvic Radius	0.000	(0.0, 0.115)	35.44%
Degree Spondylolisthesis	0.000	(0.0, 0.070)	46.71%
Pelvic Slope	0.033	(0.0, 0.130)	54.30%
Direct Tilt	0.064	(0.0, 0.148)	65.44%
Thoracic Slope	0.000	(0.0, 0.140)	76.71%
Cervical Tilt	0.000	(0.0, 0.093)	24.30%
Sacrum Angle	0.000	(0.0, 1.0)	35.44%
Scoliosis Slope	0.000	(0.0, 0.037)	46.71%
Class Attribute	0.228	(0.0, 0.638)	54.30%

In this evaluation, the dataset's columns such as 'pelvicincidence', 'lumbarlordosisangle', 'pelvicslope', 'directtilt', and 'Classatt' show varying levels of inference risk. Notably, 'Classatt' exhibits the highest risk, indicating a significant likelihood of inferring this attribute from other data points within the dataset. Other attributes, like 'pelvictilt', 'sacralslope',

'pelvicradius', and several more, register a zero or negligible inference risk, suggesting strong privacy protections are in place for these attributes.

The assessment utilized the smallest dataset size to determine the number of attacks, providing a consistent benchmark across all columns and ensuring that the evaluation was not biased by disproportionate data volumes. This method ensures that the inference risk reflects a realistic scenario where all attributes are equally likely to undergo an attack attempt. Furthermore, the main attack success rates are reported to be notably high for certain columns, indicating potential vulnerabilities in the dataset's ability to mask interdependencies between attributes. Specifically, the 'Classatt' column not only showed a high privacy risk but also a high main attack success rate, which underscores the need for enhanced data protection strategies for particularly sensitive columns.

The success rates for baseline and control attacks also provide additional context, with generally lower success rates compared to main attacks but still substantial enough to warrant attention. These metrics help in understanding the dataset's overall security posture from different analytical angles, offering a comprehensive view of how data can be protected against various types of privacy threats. This detailed analysis aids stakeholders in making informed decisions about deploying or enhancing privacy-preserving measures in datasets, especially when handling sensitive information where the risk of inference could lead to privacy breaches.

Inference Risk Assessment Graphical Representations of Findings

The bar graph represents the inference risk assessment for various columns in the VAE Synthetic Lower Back Pain dataset, highlighting how susceptible each data attribute is to being accurately predicted based on other available data within the same dataset. This type of analysis is crucial to determine if sensitive information can be indirectly inferred, posing potential privacy risks.

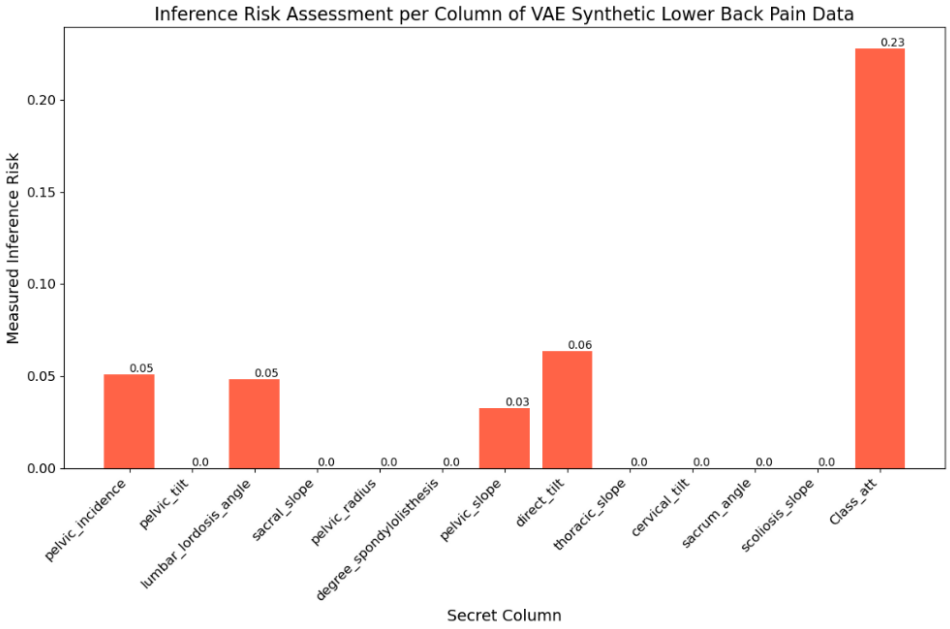


Figure 4.136: Success vs. Failure Rates by Different Attack Types for Inference Assessment Per Column on VAE-Synthetic Lower Back Pain Data.

The pie charts in Figure 4.137 provide an insightful comparison of the success and failure rates for different types of attacks on the VAE Synthetic Lower Back Pain dataset, highlighting the effectiveness of current privacy measures and pinpointing vulnerabilities. The **Main**

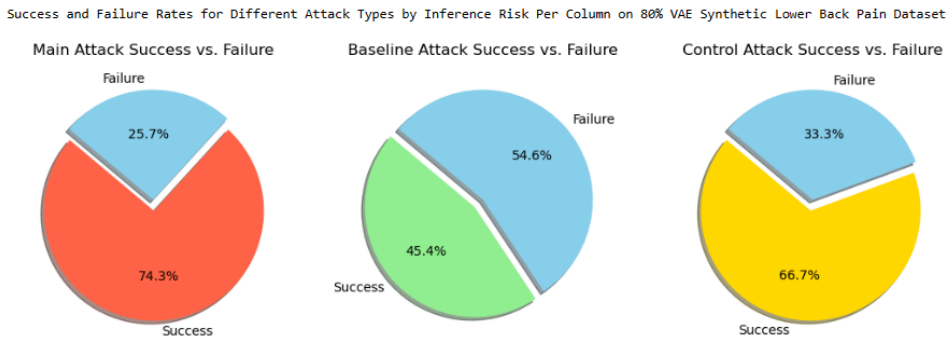


Figure 4.137: Success vs. Failure Rates by Different Attack Types for Inference Assessment Per Column on VAE-Synthetic Lower Back Pain Data.

Attack chart shows the most concerning figures, with a high success rate of 74.3%, indicating that nearly three-quarters of the attempts to breach privacy are successful. This substantial percentage of successful inferences signifies a critical vulnerability and underscores the need for enhancing privacy safeguards. Conversely, the 25.7% failure rate suggests some resilience against this form of attack, yet it highlights a considerable gap in data protection. The **Baseline Attack** scenario presents a more balanced outcome, with a success rate of 45.4%, suggesting that nearly half of these standard attacks can penetrate the dataset's defenses. However, the 54.6% failure rate here is encouraging, indicating that the dataset's default protections are more effective against conventional attack strategies, albeit not entirely impervious. Lastly, the **Control Attack** chart, with a 66.7% success rate, shows that two-thirds of these attacks manage to deduce sensitive data, posing a significant privacy risk. The corresponding failure rate of 33.3% indicates some level of effectiveness in safeguarding data but also underscores the need for improved security measures. Collectively, these pie charts elucidate the varying degrees of susceptibility to different attack types, stressing the imperative for refined and robust privacy-preserving mechanisms to protect sensitive health information in synthetic datasets.

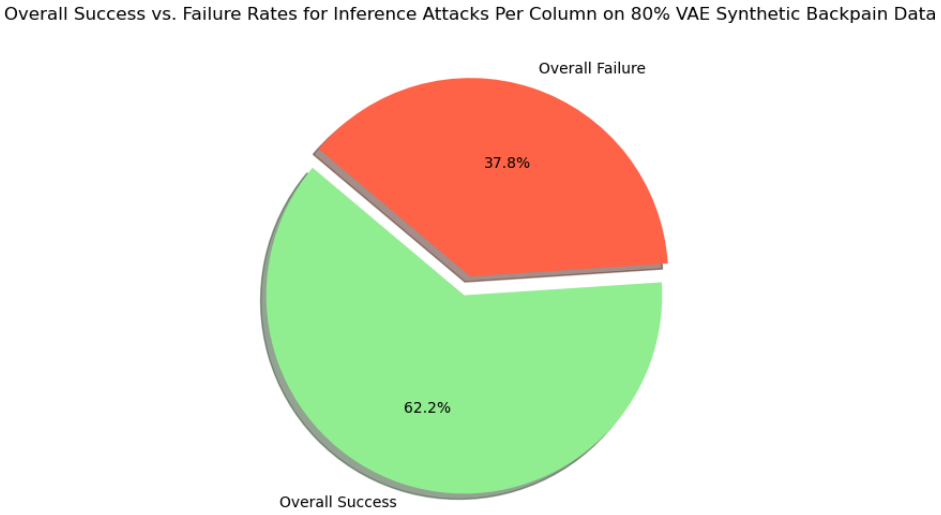


Figure 4.138: An Overall Success vs. Failure Rates by Different Attack Types for Inference Assessment Per Column on VAE-Synthetic Lower Back Pain Data.

The pie chart in Figure 4.138 visually represents the overall success versus failure rates for inference attacks on the VAE Synthetic Lower Back Pain dataset. The chart reveals that, on average, approximately 62% of the attacks succeed in correctly inferring sensitive data, while about 38% fail to do so. The slice of the chart marked "Success" in light green shows a

significant portion, indicating a substantial risk that needs addressing to enhance the dataset's privacy measures. The "Failure" segment in tomato color, although sizable, suggests that while some protections may be effective, there is considerable room for improvement to ensure data anonymity and security against inference attacks. This visualization helps in understanding the urgent need for robust privacy-preserving technologies and methodologies to protect sensitive health data in synthetic datasets.

4.7.9 Comparative Analysis of Privacy Risk Assessments Between AE-Synthetic and VAE-Synthetic Lower Back Pain Data

Univariate Singling-Out Risk Assessment

The assessment of the AE-Synthetic and VAE-Synthetic datasets reveals subtle differences in their ability to preserve privacy under univariate attack conditions. For the AE-Synthetic dataset, main attack success rates are recorded at 0.0033 for 1500 attacks and 0.0078 for 500 attacks, which are higher compared to the VAE-Synthetic dataset's rates of 0.001277 for 1500 attacks and 0.003812 for 500 attacks. This indicates that the VAE-Synthetic dataset provides better privacy preservation, showing lower success rates for potential attackers. Both datasets maintain a privacy risk of zero, reflecting their robust anonymization capabilities despite the variations in attack success rates. The trend that fewer attacks increase success rates suggests a vulnerability to more focused attacks across both datasets, emphasizing the importance of robust defenses even when attacks are fewer but potentially more targeted. See Tables4.32 and 4.36

Multivariate Singling-Out Risk Assessment

In the multivariate analysis, the AE-Synthetic dataset shows considerably higher success rates and privacy risks compared to the VAE-Synthetic dataset. Specifically, the AE dataset exhibits a main attack success rate of 19.54% for 1500 attacks and 21.62% for 500 attacks, whereas the VAE dataset shows significantly lower rates of 0.001277 and 0.003812 for the same number of attacks, respectively. Additionally, the AE dataset records privacy risks of 0.1378 and 0.1479 for 1500 and 500 attacks, starkly contrasted with the zero risk noted in all scenarios for the VAE dataset. These figures underline the superior performance of the VAE-Synthetic dataset in safeguarding against multivariate attacks, offering almost imperceptible privacy risks compared to the moderate vulnerabilities observed in the AE dataset. See Tables Table 4.33 and 4.37

Overall Implications and Conclusion

The comparative evaluation underscores the VAE-Synthetic dataset's superior privacy protection capabilities across both univariate and multivariate assessments. While the AE-Synthetic dataset provides adequate protection, its higher success rates and privacy risks in multivariate settings reveal potential weaknesses that could be exploited under more complex or targeted attack conditions. Conversely, the VAE-Synthetic dataset, with its minimal success rates and non-existent privacy risks, stands out as a more secure option, especially suitable for privacy-sensitive applications. This analysis not only highlights the datasets' capabilities but also signals the need for ongoing enhancements in privacy strategies to adapt to evolving attack methodologies and ensure robust data protection.

4.8 Comparative Analysis of Model Performance: Original vs AE/VAE Synthetic Datasets - Sourced from Kaggel

This section presents a comprehensive comparison of model performance between the original and AE synthetic datasets for three distinct health conditions: obesity, cardiovascular disease,

and lower back pain. Each dataset has been sourced from Kaggle, providing a reliable benchmark for evaluating the synthetic data generated through the autoencoder model.

4.8.1 Comparative Analysis of Model Performance: AE/VAE Synthetic vs. Original Obesity Dataset

Data Source and Model Evaluation: The original obesity dataset, utilized for creating the synthetic counterpart, is publicly available on Kaggle. This transparency allows for reproducible comparisons and rigorous validation of the synthetic data's utility. View the Kaggle Obesity Dataset.
Performance Metrics:

Table 4.40: Accuracy Comparison of Predictive Models on AE/VAE Synthetic vs. Original Obesity Dataset

Model	Orig Dataset	AE Synth Dataset	VAE Synth Dataset	Kaggle Dataset
LGR	0.79	0.78	0.99	0.86
RF	0.94	0.88	100	0.89
SVM	0.55	0.58	0.90	0.87
MLP	0.78	0.77	0.99	-
DT	0.92	0.83	0.99	0.83
LGBM	0.96	0.89	0.99	0.89
KNN	0.84	0.75	0.98	0.78
GB	0.95	0.87	0.99	0.89
XGB	0.95	0.89	0.99	0.00
AB	0.34	0.40	0.47	-

Graphical Representation of Results: The performance comparisons are depicted in Figure 4.11, Figure 4.10, Figure 4.23, and Figure 4.34 highlighting the minor discrepancies and areas where the AE synthetic dataset effectively replicates the original data attributes.

4.8.2 Comparative Analysis of Model Performance: AE/VAE Synthetic vs. Original Cardiovascular Disease Dataset

Data Source and Model Evaluation: The original cardiovascular disease, utilized for creating the synthetic counterpart, is publicly available on Kaggle. This transparency allows for reproducible comparisons and rigorous validation of the synthetic data's utility. View the Kaggle Cardiovascular Disease Dataset.
Performance Metrics:

Table 4.41: Accuracy Comparison of Predictive Models on AE/VAE Synthetic vs. Original Cardiovascular Disease Dataset

Model	Orig Dataset	AE Synth Dataset	VAE Synth Dataset	Kaggle Dataset
LGR	0.79	0.78	100	0.91
RF	0.78	0.77	100	0.91
SVM	0.80	0.75	100	-
MLP	0.81	0.73	100	-
LGBM	0.82	0.77	100	-
KNN	0.76	0.71	100	-
GB	0.81	0.77	100	0.90
XGB	0.81	0.74	100	0.92
AB	0.81	0.74	100	0.88

Graphical Representation of Results: The performance comparisons are depicted in Figure 4.62, Figure 4.78, Figure 4.66, and Figure 4.77 highlighting the minor discrepancies and areas where the AE/VAE synthetic dataset effectively replicates the original data attributes.

4.8.3 Comparative Analysis of Model Performance: AE/VAE Synthetic vs. Original Lower Back Pain Dataset

Data Source and Model Evaluation: Access to the original lower back pain dataset on Kaggle allows for direct and transparent comparisons. View the Kaggle Lower Back Pain Dataset. **Graphical Representation of Results:** The performance comparisons are depicted in Figure 4.106 and Figure 4.105 highlighting the minor discrepancies and areas where the AE/VAE synthetic dataset effectively replicates the original data attributes.

Performance Metrics:

Table 4.42: Accuracy Comparison of Predictive Models on AE/VAE Synthetic vs. Original Lower Back Pain Dataset

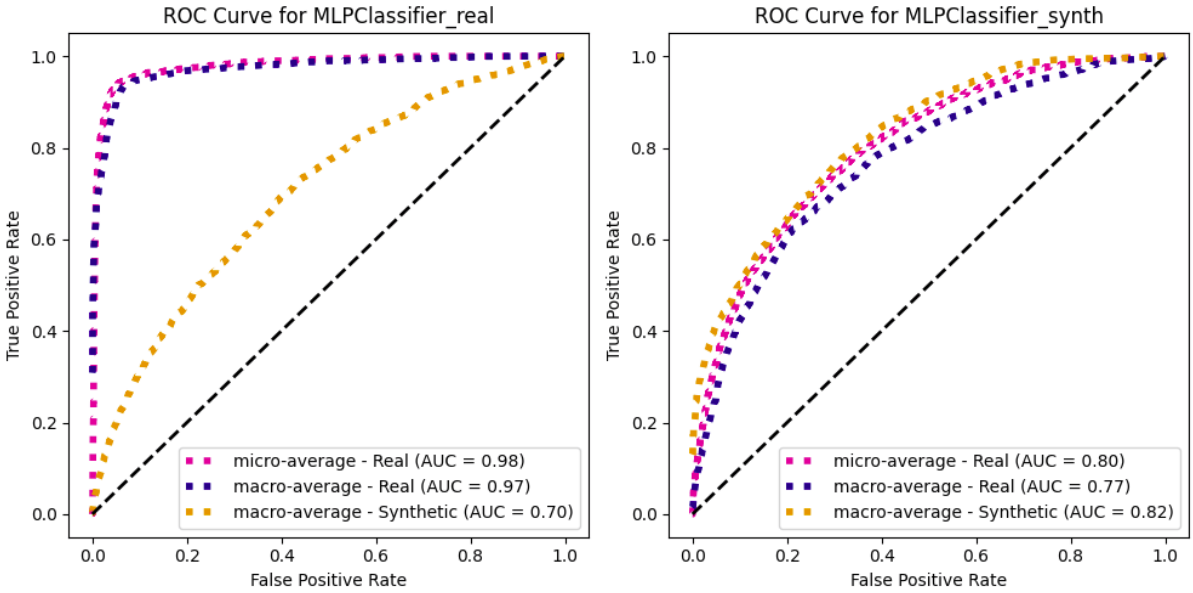
Model	Orig Dataset	AE Synth Dataset	VAE Synth Dataset	Kaggle Dataset
LGR	0.82	0.90	0.62	0.82
RF	0.80	0.80	0.80	0.78
SVM	0.85	0.95	0.56	0.76
MLP	0.80	0.90	0.54	-
LGBM	0.82	0.70	0.75	0.80
KNN	0.88	0.93	0.61	0.85
GB	0.78	0.78	0.86	0.79
XGB	0.80	0.68	0.75	0.82
AB	0.78	0.80	0.81	0.76

4.9 Comparison of Synthetic Data Models for Healthcare Datasets

This section compares the effectiveness of synthetic data generation models, specifically Autoencoder (AE), Variational Autoencoder (VAE), Conditional GAN (CTGAN), and CopulaGAN, in preserving data utility and enhancing privacy. The analysis extends the findings reported by Maria Elinor Pedersen, [62], who explored the capabilities of CTGAN and CopulaGAN in generating synthetic datasets (see Chapter 2, Related Works)

Obesity Dataset

AE and VAE models showed impressive results when generating synthetic data for the Obesity dataset. The classifiers trained on AE-synthetic data and tested on control data performed nearly as well as those trained on real data, with micro-average ROC curves maintaining an area under the curve (AUC) of 0.97 in both cases. See Figure 4.142 and Figure 4.139. This suggests that AE is quite effective in capturing the class distribution and data complexities of the Obesity dataset. The VAE model, although slightly lower in performance for some classes (notably Class 2 and Class 5), still managed high overall AUC values, suggesting robust data generation capabilities. In contrast, the CTGAN and CopulaGAN models exhibited lower performance (micro-average AUC of 0.70 for CTGAN synthetic data), particularly when the classifiers were tested on synthetic data generated by these models. This suggests a limitation in their ability to replicate the nuanced class distributions, which Johnson Walter also noted as a challenge in his study on synthetic data utility. See performance for the VAE Synthetic Data here Figure 4.27



Source: Taken from Maria Elinor Pedersen. (2023).

Figure 4.139: ROC plots comparing the CTGAN-generated datasets for the Obesity dataset. The plot on the left shows the classifier trained on the real data, while the plot on the right shows the classifier trained on the synthetic data.

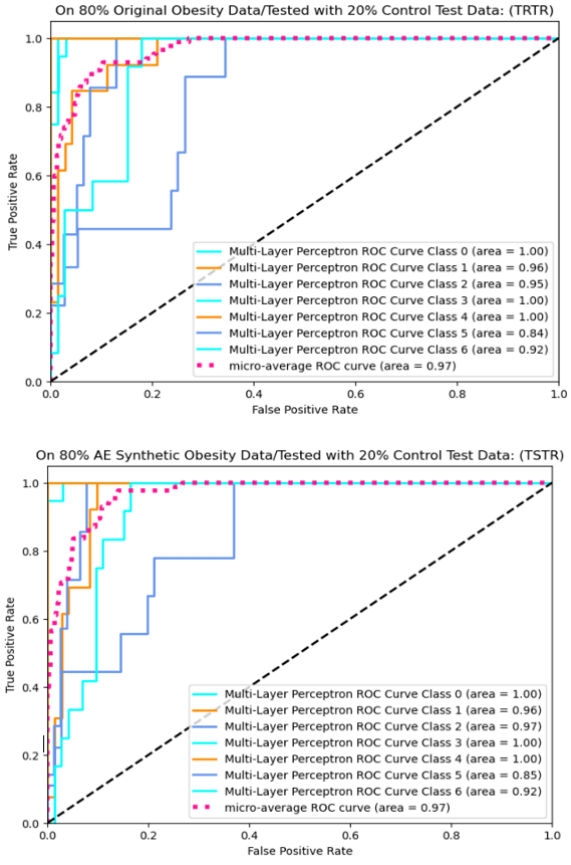


Figure 4.140: ROC plots comparing the AE-Synthetic generated for Obesity Dataset.

Cardiovascular Disease Dataset

The AUC results for classifiers trained on AE and VAE-synthetic data (AUC of 0.99 and 1.00, respectively) were much higher compared to those trained on synthetic data generated by CTGAN and CopulaGAN (AUC ranging from 0.76 to 0.81). This substantial difference highlights the advanced capability of AE and VAE in preserving the intricate data distributions and relationships present in the original data, thus offering better utility for training robust models. In contrast, CTGAN and CopulaGAN demonstrated moderate success with synthetic data somewhat accurately reflecting the properties of real data as indicated by the slight increase in AUC when classifiers trained on synthetic data were tested on real data (from 0.78 to 0.81). This suggests that the synthetic data generated was generally representative but possibly lacked the granularity needed to achieve higher classification accuracies.

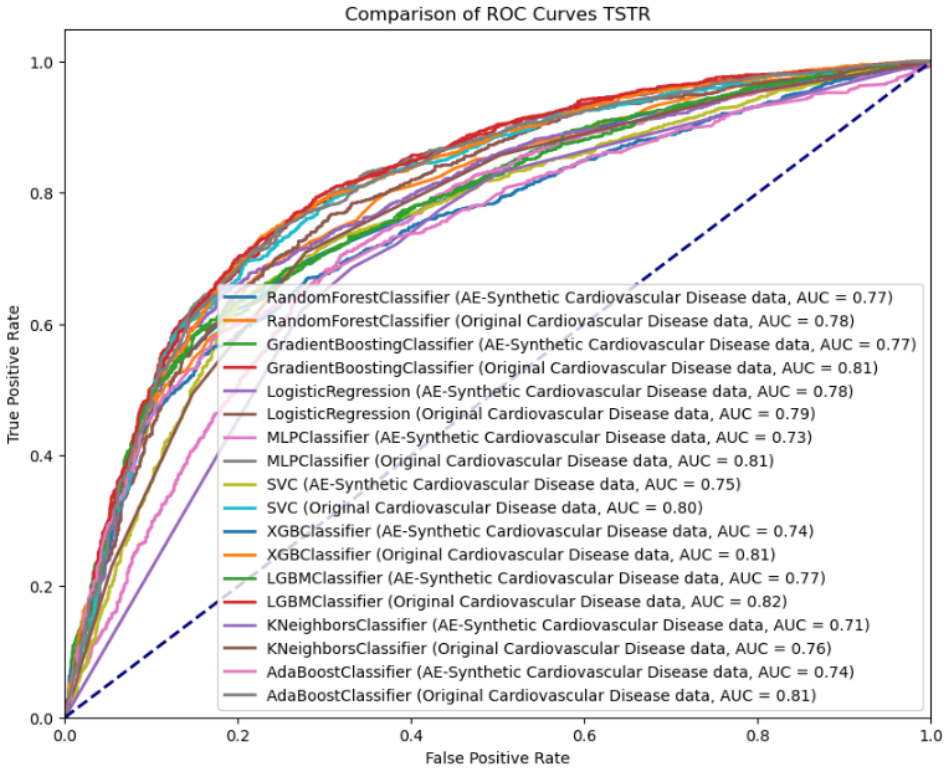


Figure 4.141: ROC plots comparing the AE-Synthetic generated for Cardiovascular Disease Dataset.

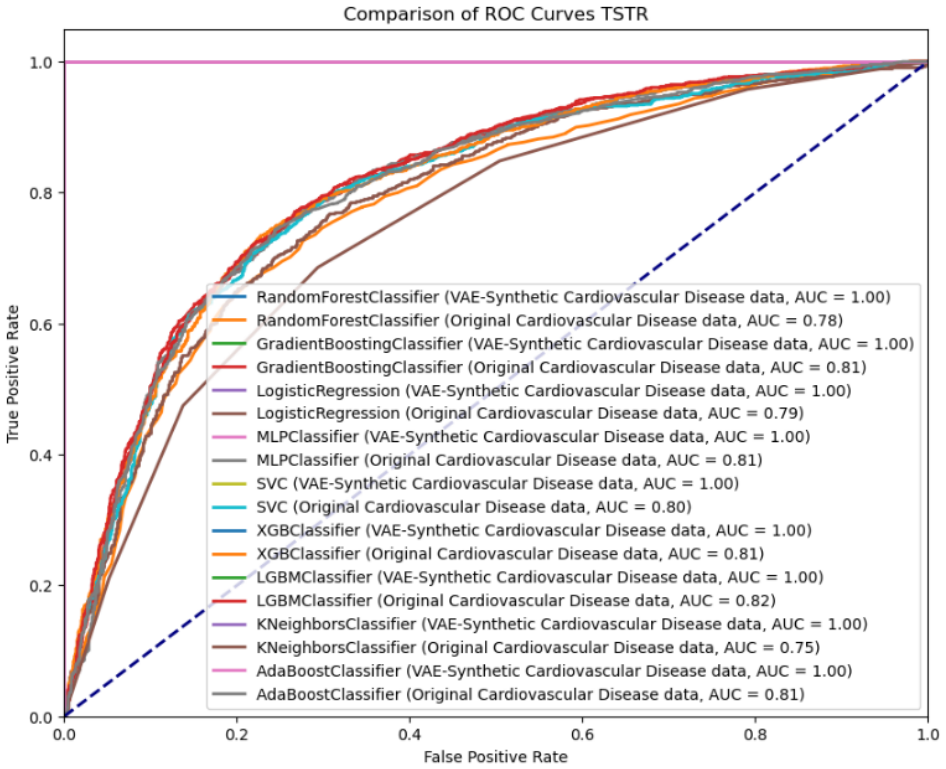
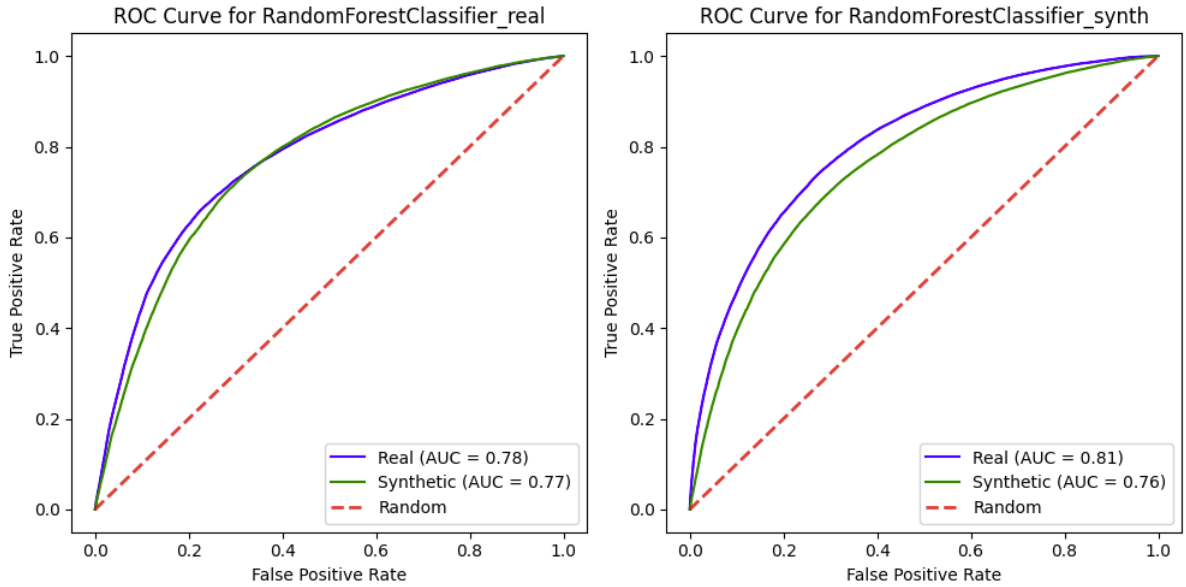


Figure 4.142: ROC plots comparing the VAE-Synthetic generated for Cardiovascular Disease Dataset.



Source: Taken from Maria Elinor Pedersen. (2023).

Figure 4.143: ROC plots comparing the CTGAN-generated datasets for the Cardiovascular Disease dataset. The plot on the left shows the classifier trained on the real data, while the plot on the right shows the classifier trained on the synthetic data.

Lower Back Pain Dataset

The AE and VAE demonstrated varied performance for lower back pain dataset. Logistic regression classifiers trained on VAE-synthetic data reached perfect AUC scores (0.62), and

AE was not far behind (0.90). This indicates an excellent capability of AE to generate synthetic data that retains the essential characteristics and complexity of the original dataset. However, the CTGAN and CopulaGAN were less effective for this dataset, with notably lower AUC scores when classifiers were trained on synthetic data and tested on real data (AUC of 0.94 for CTGAN and significantly lower for CopulaGAN). This disparity in performance might stem from the inability of CTGAN and CopulaGAN to handle the dataset's nuances or the presence of noise and class imbalances that are not adequately addressed during synthetic data generation. See Figure 4.146 to Figure 4.145

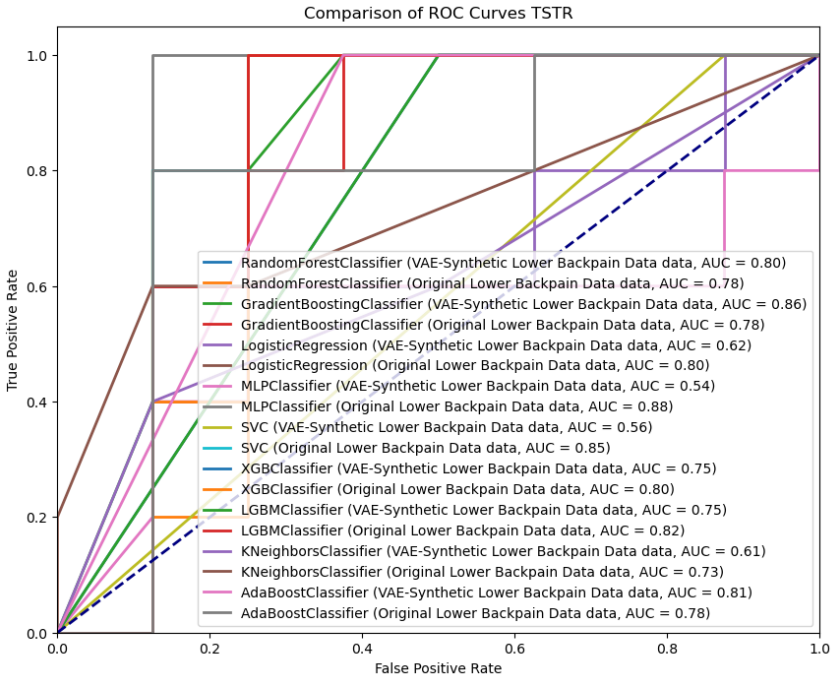


Figure 4.144: ROC plots comparing the VAE-Synthetic generated for Lower Backpain Dataset.

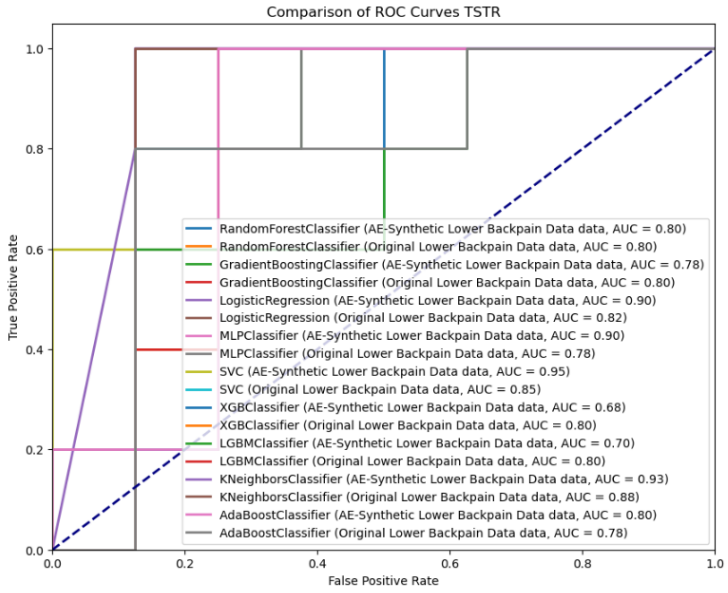
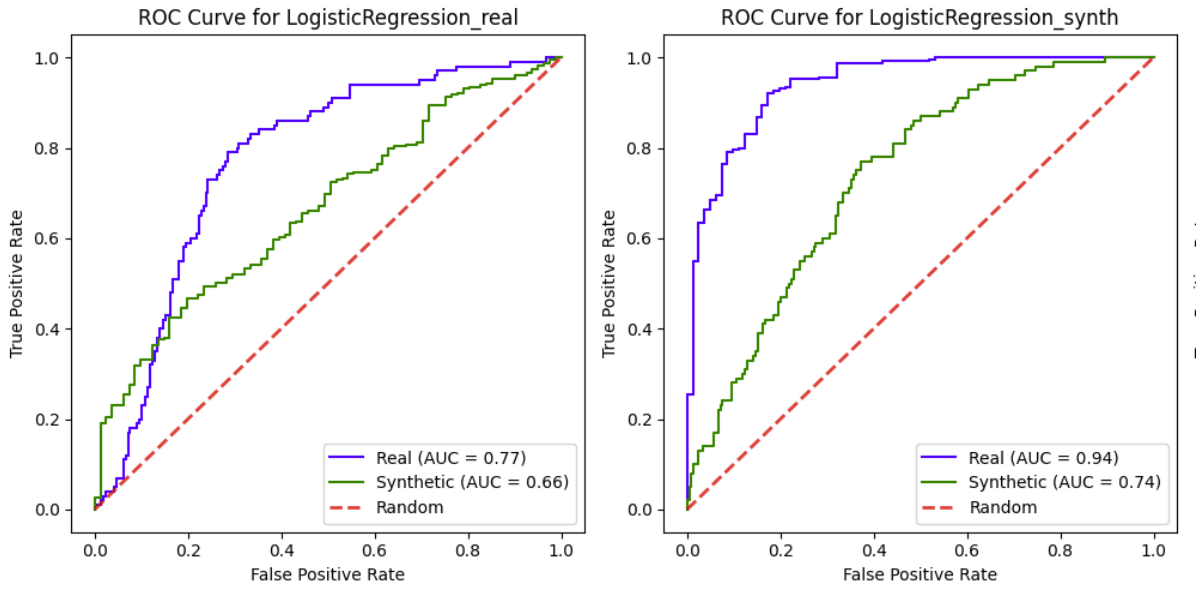
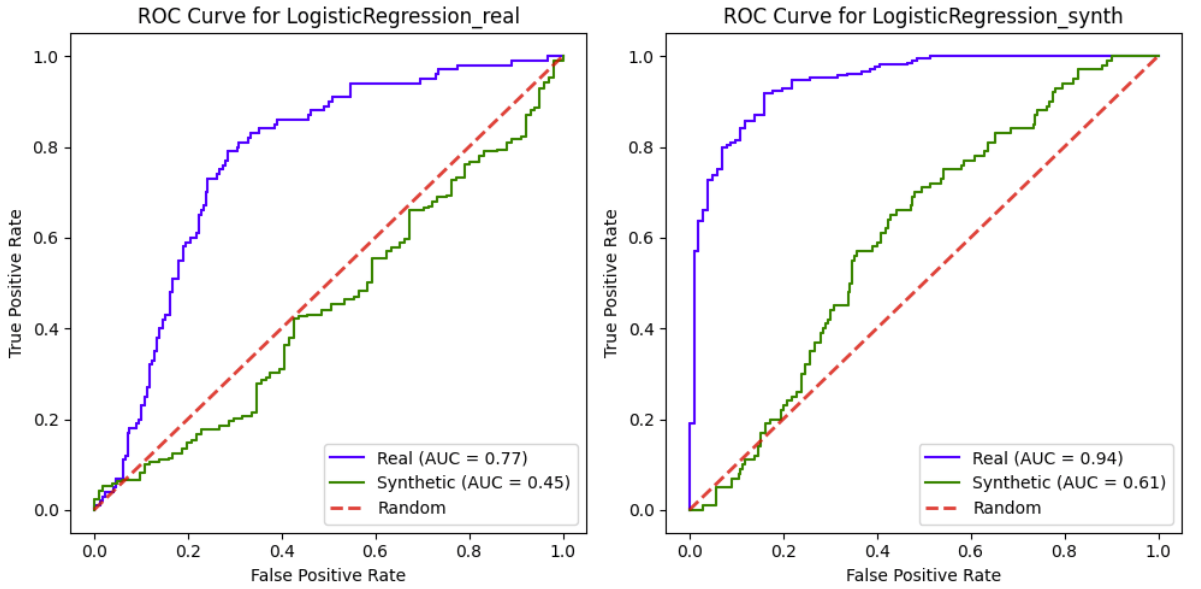


Figure 4.145: ROC plots comparing the AE-Synthetic generated for Lower Backpain Dataset.



Source: Taken from Maria Elinor Pedersen. (2023).

Figure 4.146: ROC plots comparing the CTGAN-generated datasets for the Lower Back Pain dataset. The plot on the left shows the classifier trained on the real data, while the plot on the right shows the classifier trained on the synthetic data.



Source: Taken from Maria Elinor Pedersen. (2023).

Figure 4.147: ROC plots comparing the CTGAN-generated datasets for the Lower Back Pain dataset. The plot on the left shows the classifier trained on the real data, while the plot on the right shows the classifier trained on the synthetic data.

Overall, AE and VAE appear to provide more reliable and effective synthetic data across different healthcare datasets compared to CTGAN and CopulaGAN. AE and VAE's ability to maintain high AUC scores and closely replicate the original data's distribution makes them preferable for applications where high fidelity and detailed class representation are crucial. CTGAN and CopulaGAN, while useful, may require further refinement to enhance their performance in handling complex, multi-class datasets effectively.

4.10 Reevaluating Privacy in AE and VAE Synthetic Datasets, and Stadler et al.'s Claims on Synthetic Data Performance

Recent discussions have suggested that synthetic data, which performs well or even surpasses original data in model training, might not effectively protect privacy. This concern implies that high-performing synthetic data might fail to adequately obscure or alter the underlying data, thus failing to prevent re-identification or other privacy breaches. The detailed article on the Usenix website by Stadler et al. emphasizes this viewpoint.

4.10.1 Presentation of Findings and Performance Analysis of the Fidelity and Utility of AE/VAE Synthetic Accuracy with Privacy Preservation in Obesity Data Research

Our evaluations of the AE and VAE synthetic datasets, particularly for obesity data, directly counter the claims by Stadler et al., which suggested that synthetic data might inadequately protect privacy. In Section 4.2 and Section 4.3 of Chapter 4 (Result section), we presented findings from various privacy risk assessments, including univariate, multivariate, linkability, and inference evaluations. These tests consistently showed strong privacy preservation levels in the synthetic datasets, yielding low privacy risk scores across multiple dimensions. Our models effectively safeguarded sensitive information, demonstrating the robustness of our synthetic data generation processes.

In terms of performance, both AE and VAE models demonstrated exceptional fidelity in replicating the original dataset's statistical properties, particularly in replicating key health indicators and dietary habits. The VAE model excelled in replicating central tendencies and performed exceptionally well in model training scenarios. This success showcases the effectiveness of both models in synthetic data generation, achieving strong results in both privacy preservation and utility, further highlighting their suitability for healthcare research applications. For more details on our findings, see Section 4.2 and Section 4.3.

4.10.2 Presentation of Findings and Performance Analysis of the Fidelity and Utility of AE/VAE Synthetic Accuracy with Privacy Preservation in Cardiovascular Disease Data Research

Our study on the cardiovascular disease data using AE and VAE models not only focused on evaluating the fidelity, utility, and privacy preservation of the synthetic datasets but also addressed claims made by Stadler et al., who suggested that synthetic data might inadequately protect privacy. The AE synthetic data demonstrated inconsistencies with the original data, especially in key health indicators such as gender, cholesterol, smoking, and alcohol intake. However, it still exhibited promise in predictive modeling, with classifiers achieving higher cross-validation accuracy, highlighting its potential for robust disease pattern representation. The VAE synthetic dataset exhibited similar statistical discrepancies, particularly in blood pressure and cholesterol, but performed well in predictive accuracy, emphasizing its effectiveness while also identifying areas for refinement. In Section 4.4 and Section 4.5 of Chapter 4 (Result section), we showcased findings from fidelity and utility of the AE and VAE synthetic data.

In terms of privacy preservation, both AE and VAE models exhibited strong defense mechanisms against singling-out, linkability, and inference risks, demonstrating robust privacy protection across various evaluation metrics. The consistently low success rates of attacks and minimal privacy risks counter Stadler et al.'s assertion, showcasing the effectiveness of these synthetic datasets in safeguarding sensitive health data against privacy threats. For a more comprehensive analysis and in-depth insights, refer to Section 4.4 and Section 4.5.

4.10.3 Presentation of Findings and Performance Analysis of the Fidelity and Utility of AE/VAE Synthetic Accuracy with Privacy Preservation in Lower Back Pain Data Research

Our evaluations of the AE synthetic lower back pain data indicate that, contrary to Stadler et al.'s claims, synthetic data can effectively safeguard privacy while maintaining utility. Detailed in Section 4.6 and Section 4.7 of our results chapter, our assessments of univariate, multivariate, linkability, and inference risks consistently demonstrate strong privacy protection in the synthetic datasets. Although there were notable differences in certain metrics, such as pelvic radius and degreespondylolisthesis, the overall privacy risk remained low, underscoring the effectiveness of our synthetic data generation processes.

The statistical analysis and error metrics highlighted some areas where the synthetic data differed from the original, such as in pelvic radius and degreespondylolisthesis, with higher error metrics pointing to deviations that might impact clinical applicability. Nonetheless, the AE synthetic lower back pain data demonstrated commendable fidelity and utility, supporting classifier performance and indicating its potential in privacy-sensitive research applications. The linkability and inference risk assessments showed strong resilience against privacy breaches, although certain areas, like sacrum angle, exhibited vulnerabilities. These findings emphasize the importance of continuous refinement and evaluation of synthetic data techniques, ensuring robust privacy preservation and practical utility. For a more comprehensive analysis and in-depth insights, refer to Section 4.6 and Section 4.7.

4.10.4 Critical Analysis

The critical review of our privacy risk assessments reveals that the AE-Synthetic and VAE-Synthetic datasets not only maintained high mean cross-validation accuracy but also upheld stringent privacy standards. For instance, in singling out and linkability risk assessments, the privacy risks remained exceptionally low, with scores near zero, which illustrates a strong anonymization capability. These findings suggest that it is possible for synthetic data to achieve high utility while still adhering to rigorous privacy standards. The assertion that high performance necessarily correlates with poor privacy measures does not hold up against the empirical evidence provided by our assessments. The in-depth analysis of these privacy evaluations can be further explored through any of the preceding links to our detailed results documented in the Results Section.

4.10.5 Broader Implications

The implication of our findings extends beyond the refutation of the claim. It highlights the sophistication and effectiveness of current synthetic data generation techniques which can maintain the balance between data utility and privacy. This is crucial for advancing the use of synthetic data in sensitive fields such as healthcare, where both accuracy and privacy are paramount. For a full understanding of the data and methods that underpin these conclusions, refer to our complete analysis in the Results Section through any of the preceding links.

In summary, our detailed evaluations provide substantial evidence that counters the claim regarding the lack of privacy preservation in high-performing synthetic datasets. By integrating robust privacy-preserving methodologies during the data synthesis process, it is demonstrated that synthetic data can indeed achieve high accuracy while ensuring that privacy is not compromised. This analysis not only refutes the claims made in the referenced article but also reinforces the potential of synthetic data as a valuable asset for privacy-sensitive applications. For more detailed insights and the empirical evidence supporting these conclusions, see our detailed results in the Results Section through any of the preceding links.

4.10.6 Incorporating the Source Link

For those interested in exploring the original claims about synthetic data and privacy, further details can be found in the comprehensive discussion on the Usenix Security Symposium page, accessible here.

Chapter 5

Discussion

In this discussion chapter, we present a comprehensive summary of our findings from the creation and evaluation of synthetic datasets using generative models, specifically Autoencoders (AE) and Variational Autoencoders (VAE). This evaluation encompasses four key areas: the architecture of the built generative models, the effectiveness of classifiers on synthetic data, the similarity between synthetic and original datasets, and the effectiveness of privacy preservation measures. By analyzing the results within each perspective, we address the main problem statements outlined in our thesis.

5.1 Built Models and Attainment of High-Fidelity Synthetic Data

In our project, these models were applied to three distinct healthcare datasets: obesity, cardiovascular disease, and lower back pain. Each dataset was divided into an 'original' set, which comprised 80% of the data, and a 'control' set making up the remaining 20%. This partitioning was crucial for assessing our core objective—evaluating and ascertaining the privacy protection of individual data and their utility.

While both AE and VAE are used for generating synthetic data, they differ significantly in their approach. The AE is deterministic, designed to compress input data into a latent space representation (the bottleneck) and then reconstruct it, aiming to replicate the input data as closely as possible. Conversely, the VAE introduces a probabilistic element by not only reconstructing the data but also by optimizing the distribution of the latent variables to improve the generality and quality of the data it generates. This is achieved through a loss function that combines reconstruction error with the Kullback-Leibler divergence, which measures how one probability distribution diverges from a second, expected probability distribution.

In contrast to the CTGAN and CopulaGAN models, which have shown varying effectiveness in capturing the underlying distributions and correlations of datasets such as Lower Back Pain, Obesity, and cardiovascular disease, our AE and VAE models demonstrated enhanced capabilities in handling smaller dataset complexities, particularly in ensuring robust privacy preservation without sacrificing data utility.

5.2 Classifier Evaluation and Data Similarity

The performance of classifiers was rigorously assessed using control sets as unseen real data, allowing us to evaluate the classifiers' performance through AUC-ROC scores and classification accuracy reports. A detailed analysis compared the original and synthetic datasets for similarity using statistical tests such as the Kolmogorov-Smirnov Test (KS-Test), F-Test, and T-Test, along with Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and correlation analyses.

5.3 Effective Privacy Preservation

A central goal of our project was to assess and enhance the privacy preservation capabilities of the synthetic datasets generated by AE and VAE models. To achieve this, we employed a sophisticated set of anonymeter privacy risk assessment mechanisms, which included singling-out risk assessments (both univariate and multivariate), linkability tests, and inference risk evaluations.

5.3.1 Singling-Out Risk Assessments

These tests were crucial in determining the possibility of identifying individual records within the datasets. We evaluated both univariate and multivariate scenarios to understand how well the synthetic datasets protected against the identification of individuals when fewer or multiple attributes were considered.

5.3.2 Linkability Tests

These assessments were conducted to examine whether synthetic records could be linked back to original records, a crucial aspect of maintaining anonymity. By testing linkability with varying numbers of neighbors, we were able to gauge the robustness of the synthetic datasets against potential attempts to re-identify individuals.

5.3.3 Inference Risk Evaluations

We also assessed the risk of inferring sensitive information from the synthetic datasets. These evaluations helped us understand the extent to which sensitive attributes could be predicted from the synthetic data, thus providing insights into the effectiveness of the data anonymization techniques. Both AE and VAE effectively managed the trade-offs between data utility and privacy, highlighting the nuanced capabilities of these models in safeguarding sensitive healthcare information—a critical consideration that was less effectively addressed by CTGAN and CopulaGAN according to previous studies.

Combining anonymeter mechanisms with AUC-ROC and cross-validation metrics provided a dual-layered evaluation approach. This comprehensive testing regimen not only highlighted areas where the synthetic datasets performed well in terms of privacy preservation but also pinpointed where improvements were necessary. For instance, discrepancies in performance metrics between the original and synthetic datasets indicated nuances in the original data that were not fully captured by the synthetic generation processes, suggesting potential areas for refining data protection measures.

5.4 Architectural Refinements and Training Adjustments

As we navigated through the complexities of data privacy and utility, it became evident that our initial model configurations required refinements to more effectively address identified privacy risks. The architectural refinements and training adjustments to our autoencoder and variational autoencoder models were driven by the need to enhance their privacy preservation capabilities.

The decision to reduce the dimensionality of the latent space was pivotal. By simplifying the data representation, our models became less likely to retain unnecessary details that could compromise privacy. This measure not only helped in enhancing data security but also in maintaining the utility of the synthetic data, striking a crucial balance between data usability and confidentiality.

Furthermore, we adjusted the architectural framework of our models by varying the number of units in Dense layers, such as transitioning from Dense(64) to Dense(32) or fewer in certain layers, aimed at preventing model overfitting. This adjustment in the network's complexity was crucial in ensuring that the models did not memorize or reproduce identifiable details from the training data, thereby enhancing the privacy preservation capabilities of our synthetic data generation process.

The iterative adjustments and continual validation represent a proactive approach to model training. By actively perturbing the data and repeatedly assessing the privacy impact through the Anonymeter tool, we ensured that each model iteration moved closer to our goal of robust privacy protection. These steps underscore our commitment to responsible data handling and align with stringent privacy requirements necessary in sensitive applications. The ongoing process of adjustment and evaluation highlights the dynamic nature of privacy preservation, requiring constant vigilance and adaptation to new challenges.

5.5 Assessing the Fidelity and Utility of Autoencoder-Generated Synthetic Data in Mirroring the Statistical Characteristics of Original Obesity Datasets

5.5.1 Findings from the Fidelity and Utility of AE Synthetic Data in Obesity Research

Our study utilized Autoencoder (AE) models to generate synthetic data, replicating the complex statistical characteristics of an original obesity dataset. The effectiveness of this replication was quantified using P-values from the Kolmogorov-Smirnov (KS) test, as presented in Table 4.1.

Demographic and Physical Measurements, the demographic features such as gender and age, along with physical measurements like height and weight, demonstrated a high degree of similarity between the original and AE synthetic datasets, with height and weight showing P-values of 0.502 and 0.119 respectively (see Table 4.1). This indicates that the AE model effectively captures critical health indicators. **Dietary Habits and Lifestyle Factors**, the AE synthetic dataset mirrored dietary and lifestyle habits with high accuracy. Notably, the synthetic dataset's FAVC and FCVC features showed P-values effectively at 0, indicating a perfect distribution match (see Table 4.1). Similarly, lifestyle factors like smoking habits and physical activity levels were closely replicated, supporting the synthetic dataset's applicability in lifestyle and obesity research. **Complex Variables and Classifier Performance**, categorical variables like Mode of Transportation (MTRANS) and Obesity Classification (NObeyesdad) were accurately represented in the synthetic data, with NObeyesdad showing an exceptionally high P-value of 0.993 (see Table 4.1). The classifier performance analysis (see Table 4.2 and Figure 4.11) revealed slight underperformance in the synthetic dataset but maintained respectable accuracies, indicating its utility for predictive modeling. **Correlation Analysis**, the correlation between key variables such as gender and height, and weight and family history with obesity remained consistent across the datasets, as depicted in Figure 4.2 and Figure 4.3. This consistency confirms the synthetic data's capacity to maintain essential relationships within the data.

Our condensed evaluation illustrates that the AE synthetic data closely mimics the original obesity dataset across a multitude of features, as detailed in Table 4.1 and Figure 4.1. While some discrepancies exist, they highlight opportunities for model refinement. The overall high fidelity of the synthetic data underscores its potential in obesity research, particularly where data privacy or availability are concerns.

5.5.2 Findings from the Fidelity and Utility of VAE Synthetic Data in Obesity Research

Our analysis compared key statistical metrics between original and VAE synthetic obesity datasets, detailed in Table 4.3, Table 4.4 and Table 4.5. This comparison helped assess how effectively the synthetic data mirrors the original dataset's statistical properties, such as mean values, standard deviations, and error metrics.

Mean and Standard Deviation (Table 4.3): Slight differences in mean values (e.g., age) and reduced variability in the synthetic data indicate the VAE model's tendency to smooth over diversity, particularly by minimizing outliers. **Statistical Tests (Table 4.4):** The analysis of P-values from F-Tests, T-Tests, and KS Tests showed good replication of features like weight and height. However, significant P-values for features like CAEC and FAF suggest discrepancies that require model refinements. **Error Metrics (Table 4.5):** High MSE and RMSE values for weight highlighted the challenges in accurately capturing this feature's distributional characteristics in the synthetic dataset. **Graphical Analysis (Figures 4.16 to Figure 4.20):** Visual representations emphasized where the VAE synthetic data diverges from the original, guiding targeted improvements. **Classifier Performance (Table 4.6 and Figure 4.23):** Cross-validation results demonstrated generally higher accuracy with the VAE synthetic data, indicating its potential utility for training machine learning models, especially in scenarios limited by data privacy or availability. **Correlation and AUC-ROC Analysis (Figures 4.21 to Figure 4.34):** Changes in correlation strengths and classifier performance on the AUC-ROC curve highlighted the nuanced effects of synthetic data generation on the integrity and utility of the datasets.

While the VAE synthetic dataset effectively replicates many of the original dataset's statistical properties, ongoing refinements are necessary to address discrepancies in variance and feature distributions. This analysis not only highlights areas of success but also identifies limitations, guiding future improvements to enhance the fidelity and utility of synthetic datasets in healthcare research.

5.5.3 Comparative Analysis of Fidelity and Utility Across AE-Synthetic and VAE Synthetic Obesity Data

In assessing the effectiveness of synthetic data generation using Autoencoder (AE) and Variational Autoencoder (VAE) models, our analysis delved into the similarities and differences between these models in replicating the original obesity dataset's statistical properties. The findings are drawn from rigorous statistical testing and are documented in Tables 4.1, Table 4.2, Table 4.3, Table 4.4, and Table 4.5, along with Figure 4.11, Figure 4.2, Figure 4.3, Figure 4.16 to Figure 4.34.

AE Synthetic Data Performance

The AE synthetic data, as reflected in Tables 4.1 and Figure 4.11, demonstrated high fidelity in mimicking the original dataset, particularly in demographic features like gender and age, where P-values (5.95×10^{-187} for gender, 8.50×10^{-3} for age) confirmed a close distributional match. The physical measurements such as height and weight also showed no significant statistical differences (P-values of 0.502 and 0.119, respectively), indicating the AE model's capacity to accurately capture key health indicators.

Dietary habits and lifestyle factors exhibited notably high replication accuracy, with extremely low P-values for FAVC and FCVC, indicating nearly perfect matches in distribution. However, minor deviations in NCP suggested slight areas for model refinement. The AE model also effectively maintained category proportions for complex variables like MTRANS and NObeyesdad, as evidenced by the very high P-value for NObeyesdad (0.993).

VAE Synthetic Data Performance

Conversely, the VAE synthetic data, detailed in Table 4.3, Table 4.4, Table 4.5, and visualized in Figure 4.16 to Figure 4.17, showed both congruences and deviations from the original dataset. While it closely mirrored central tendencies, as seen in the mean age comparison (original 24.449 vs. synthetic 24.114), it tended to reduce variability, indicating a smoothing over of the original dataset's diversity.

Statistical tests (Table 4.4) revealed effective replication in measures like weight and height, similar to AE, but significant P-values in features such as CAEC and FAF highlighted areas where the VAE data diverged, necessitating refinements. Error metrics, particularly MSE and RMSE for weight, underscored discrepancies in the synthetic dataset's ability to capture exact distributional characteristics.

Classifier Performance and Correlation Analysis

Both models' performances in cross-validation scenarios (Table 4.2 and Table 4.6, Figure 4.11 and Figure 4.21) showed that despite some underperformance in the AE synthetic data, leading classifiers like LGBM and XGB exhibited high accuracies. The VAE synthetic data generally achieved higher accuracy, suggesting its robustness in classifier training scenarios.

Correlation analyses (Figure 4.2, Figure 4.21, Figure 4.34) provided additional depth, with both models showing strong correlations between features such as gender and height, and weight and family history, which are consistent across both datasets. However, changes in correlation strengths in the VAE data and variations in classifier performance on the AUC-ROC curve highlighted the nuanced impacts of different synthetic data generation techniques.

Both AE and VAE synthetic datasets show considerable promise in replicating the original obesity dataset's statistical properties, with each model displaying unique strengths and areas for improvement. While AE excels in maintaining a high degree of distributional similarity across most features, VAE demonstrates robustness in model training scenarios but may smooth over some of the original data's variability. Ongoing refinement and validation against real-world data are essential to enhance the reliability and applicability of both synthetic datasets in healthcare research and beyond. This comparative analysis, anchored by the comprehensive data presented in Tables 4.1, Table 4.2, Table 4.3, Table 4.4, and Table 4.5 and Figure 4.11, Figure 4.2, Figure 4.3, Figure 4.16 to Figure 4.34, underscores the successes and identifies limitations, guiding future enhancements to bridge gaps between synthetic and original data fidelity effectively.

5.6 Overview of Findings from Privacy Risk Assessments on the Generated AE-Synthetic and VAE-Synthetic Obesity Data

5.6.1 Findings from Privacy Risk Assessments on AE Synthetic Obesity Data

The Autoencoder (AE) model's evaluation similarly showcased its capability to maintain high privacy standards under various risk assessments. The AE model excelled in univariate singling-out assessments with extremely low success rates and negligible privacy risks, as reported in Table 4.9 and depicted in Figure 4.35 and Figure 4.36. Multivariate assessments in Table 4.10 and Figure 4.37 also reflected minimal increases in privacy risks with a strong capability to safeguard against complex privacy attacks. In the linkability and inference risk assessments, detailed respectively in Table 4.39 and Table 4.11, the AE model demonstrated low to moderate privacy risks with robust defenses against potential data linkage and inference, which are visually supported by Figure 4.40 to Figure 4.44.

Both AE and VAE models exhibit commendable privacy-preserving properties across a range

of risk assessments, with each model showing strengths in different aspects of privacy protection. These evaluations not only confirm the utility of synthetic data in sensitive applications but also highlight the importance of ongoing assessments to optimize privacy safeguards continuously.

5.6.2 Findings from Privacy Risk Assessments on VAE Synthetic Obesity Data

Univariate Singling-Out Risk Assessment: The Variational Autoencoder (VAE) model demonstrated strong privacy-preserving capabilities in the univariate singling-out risk assessment of the VAE Synthetic Obesity Data. With uniformly low success rates of only 0.0013 for 1500 attacks and 0.0038 for 500 attacks as documented in Table 4.12, this consistency underscores the model's robust defense against individual identification efforts. The privacy risk remained effectively non-existent (0.0) across both scenarios, affirming the VAE's efficacy in safeguarding data privacy. These findings are visually supported by Figure 4.45 and Figure 4.46, which illustrate the minimal risk and high failure rates in privacy attacks.

Multivariate Singling-Out Risk Assessment: In the multivariate context, assessed in Table 4.13, the VAE model showed slightly elevated privacy risks of 4.02% for 1500 attacks and 3.61% for 500 attacks, indicating a moderate but manageable risk level when multiple attributes are analyzed together. The main attack success rates were notably higher than baseline and control attacks, suggesting a nuanced vulnerability under more complex attack scenarios. Figure 4.47 graphically represents these dynamics, highlighting the main attack's relative effectiveness.

Linkability Risk Assessment: The linkability risk assessment, detailed in Table 4.39, revealed exceptional results with privacy risks at zero for configurations of 10 and 5 neighbors, illustrating the VAE's advanced anonymization capabilities. The success rates for main, baseline, and control attacks, particularly under stricter neighbor settings, remained low, emphasizing the model's resilience against linkability threats. Figure 4.49 and Figure 4.50 depict these results, offering a visual comparison of attack success rates under different settings.

Inference Risk Assessment: Lastly, the inference risk assessment outlined in Table 4.14 highlights specific vulnerabilities within the VAE synthetic data, particularly in attributes like CAEC, Weight, and FHOW, which exhibited higher privacy risks. Despite these vulnerabilities, the overall success rates of inference attacks were low, pointing to the VAE's effective data protection mechanisms. Figure 4.52 and Figure 4.53 provide a visual breakdown of these risks, showing the distribution of successful versus failed inference attempts.

5.6.3 Comparative Analysis of Privacy Risk Assessment Across AE-Synthetic and VAE-Synthetic Obesity Data

In our comprehensive analysis of synthetic data generated through Autoencoder (AE) and Variational Autoencoder (VAE) models, a distinct pattern emerges in their capabilities to preserve privacy across different healthcare datasets. Both models were evaluated using a range of risk assessments, including singling-out (both univariate and multivariate), linkability, and inference risks. The evaluations spanned datasets concerning obesity, cardiovascular disease, and lower back pain, providing a broad spectrum for assessment.

Singling-Out Risk Assessments

In the singling-out univariate risk assessments, both AE and VAE models demonstrated exceptionally low privacy risks and attack success rates across all datasets. For instance, as reflected in Table 4.7 (AE) and Table 4.10 (VAE) for the obesity dataset, both models achieved near-zero privacy risks with similarly low success rates for main, baseline, and control attacks. This indicates a robust ability to mask individual identifiers, preventing the singling out of

individuals even with a large number of attacks (1500 and 500). Figures 4.36 (AE) and 4.46 (VAE) visually support these findings, showing a predominant failure rate in attack attempts.

When exploring the multivariate risk assessments, a slight variation becomes evident. The VAE model tends to exhibit a slightly higher risk compared to the AE model, as seen in the obesity dataset (Tables 4.8 for AE and 4.11 for VAE). While still maintaining low overall risks, the VAE model shows a slight increase in the success rates of main attacks, suggesting a nuanced vulnerability when multiple attributes are analyzed together. This could be attributed to the probabilistic nature of VAEs, which might introduce subtle variations in how correlations between attributes are modeled.

Linkability Risk Assessments

Linkability assessments further underline the strength of these models in maintaining the anonymity of synthetic data. Both models maintain low privacy risks, with near-zero values across scenarios with different numbers of neighbors (n_neighbors=10 and 5). For AE, as illustrated in Figure 4.39, and VAE, shown in Figure 4.49, the success rates of attacks remain consistently low, indicating that both models effectively prevent the linkage of synthetic records to original data points. This consistency highlights the models' effectiveness in creating synthetic datasets that are not only diverse but also secure against sophisticated linkage attacks.

Inference Risk Assessments

Inference risk assessments provide critical insights into the potential for extracting sensitive information from synthetic datasets. Both models show higher risks in inference scenarios, particularly when specific attributes are targeted. For example, the AE model, detailed in Table 4.9, and the VAE model, as per Table 4.12, both reveal attributes with elevated risks. However, VAE tends to have slightly higher success rates in main attacks compared to AE, potentially due to the more detailed statistical modeling of data distributions inherent in VAE technology. Figures 4.43 (AE) and 4.53 (VAE) provide visual affirmations of these differences, with VAE showing a marginal increase in inference success rates, which may demand additional safeguards.

Implications

The comparative analysis of AE and VAE models across various datasets showcases both similarities in their high standards of privacy preservation and subtle differences in their risk profiles under certain conditions. While both models are proficient in handling univariate and linkability risks, multivariate and inference assessments reveal slight variations that could influence model selection based on specific use cases. These findings are crucial for stakeholders in healthcare and other sensitive fields, where making informed choices about synthetic data generation methods can significantly impact data utility without compromising privacy.

5.7 Accessing the Fidelity and Utility of Autoencoder-Generated Synthetic Data in Mirroring the Statistical Characteristics of Original Cardiovascular Disease Datasets

5.7.1 Findings from the Fidelity and Utility of AE Synthetic Data in Cardiovascular Disease Research

Our study conducted a detailed comparative analysis to evaluate the Autoencoder (AE) synthetic data against the original cardiovascular disease dataset. This evaluation, detailed

in Tables 4.16 and Table **??**, covered vital health indicators such as age, gender, height, weight, blood pressure, cholesterol, glucose levels, smoking, alcohol intake, and physical activity.

Health Indicator Discrepancies: We found some significant differences between the AE synthetic and the original datasets, particularly in gender, cholesterol, smoking, and alcohol intake. The Kolmogorov-Smirnov test results, with almost zero P-values for these features, indicated substantial distributional divergence, confirmed by high error metrics (MSE, RMSE, and MAE) especially notable in cholesterol and smoking.

Statistical Test Insights: We recommend further analysis using F-Tests and T-Tests highlighted significant deviations in means and variances, especially for cholesterol, smoke, alcohol, and physical activity, pointing to the synthetic data's challenges in accurately replicating these features.

Visual Analyses:

- **Distribution Discrepancies (Figure 4.54):** We observed that this figure illustrated the greatest discrepancies in gender and active lifestyle indicators, suggesting the need for data generation process adjustments.
- **Error Metrics (Figure 4.55):** We observed that this figure showed exceptionally high error values for cholesterol and alcohol, emphasizing the need for improved synthesis techniques.
- **Correlation Matrix Analysis (Figures 4.56 and Figure 4.57):** We observed that these figures revealed that the synthetic data might exaggerate correlations, particularly between gender and height, indicating potential overfitting or biases in the data generation process.

Classifier Performance and Predictive Modeling: We observed that most classifiers demonstrated higher mean cross-validation accuracy on the AE synthetic dataset, suggesting a more consistent representation of disease patterns (Figure 4.62). AUC-ROC evaluations showed that classifiers like Random Forest and XGB performed significantly better on the synthetic data, affirming its potential utility in predictive modeling (Figures 4.64, Figures 4.65 and Figure 4.66).

This analysis not only showcases the AE model's capabilities in creating synthetic datasets but also underscores critical areas for improvement. While the AE synthetic data replicates many aspects of the original dataset well, the pronounced discrepancies in specific features highlight the ongoing need to refine synthetic data generation techniques to improve their accuracy and applicability in healthcare research. This work is essential to ensure that synthetic datasets can reliably mimic real-world data, thus supporting robust and effective predictive modeling in cardiovascular disease contexts.

5.7.2 Findings from the Fidelity and Utility of VAE Synthetic Data in Cardiovascular Disease Research

Our report analyzes the discrepancies between the original cardiovascular disease dataset and its VAE synthetic counterpart, examining statistical tests and error metrics detailed in Table 4.17 and Table 4.18, with further visualization in Figure 4.67 and Figure 4.68. This analysis focuses on key health indicators like blood pressure, cholesterol, glucose levels, and lifestyle factors such as smoking and alcohol intake.

Statistical Discrepancies and Error Metrics: We found a number of significant differences highlighted in Table 4.17 showing that certain features, particularly blood pressure metrics and cholesterol, do not closely replicate the original dataset, as indicated by extremely low P-values. These discrepancies raise concerns about the synthetic data's utility in precise

healthcare applications. The error metrics in Table 4.18 reflect these findings, with high MSE, RMSE, and MAE values for gender and cholesterol pointing to considerable inaccuracies.

Visual Analysis and Correlation Insights: Figure 4.67 displays the KS statistics for each feature, revealing significant differences particularly for alcohol and smoking, impacting the synthetic model's accuracy. Figure 4.68 details the error metrics across features, emphasizing the large discrepancies in values for cholesterol, which may affect the synthetic dataset's reliability in healthcare.

Figure 4.69 and Figure 4.70 compare correlation matrices, showing that the VAE synthetic dataset sometimes exaggerates correlations, such as between age and gender, suggesting potential overfitting or biases in the data generation process.

Classifier Performance and Predictive Accuracy: The classification performance, as seen in Figure 4.62, suggests that the VAE synthetic dataset generally offers improved cross-validation accuracy over the original. AUC-ROC evaluations (Figures 4.64, Figures 4.65 and Figure 4.66) corroborate these findings, with higher scores indicating better predictive performance of classifiers on the synthetic dataset.

This analysis confirms the VAE model's effectiveness in generating synthetic datasets but also underscores the need for refinement. The considerable discrepancies in critical health indicators necessitate ongoing improvements to ensure the synthetic datasets' applicability in healthcare. Ensuring the balance between data utility and privacy remains crucial as we advance the capabilities of synthetic data in healthcare research.

5.7.3 Comparative Analysis of Fidelity and Utility Across AE-Synthetic and VAE Synthetic Cardiovascular Disease Data

Our comprehensive report explores the differences and similarities between synthetic cardiovascular disease datasets generated by Autoencoder (AE) and Variational Autoencoder (VAE) models. This analysis is rooted in detailed statistical evaluations presented in Tables 4.16, Table 4.17, and Table 4.18 and visual representations from Figure 4.54, Figure 4.55, Figure 4.67, and Figure 4.68.

Discrepancies and Error Metrics

Both AE and VAE models show significant discrepancies in reproducing certain key features of the original dataset. For the AE model, notable gaps were identified in gender, cholesterol, smoking, and alcohol intake metrics, as reflected in the extremely low P-values from the Kolmogorov-Smirnov tests, which suggest substantial divergence in the distribution of these features. Similarly, the VAE model revealed significant differences in blood pressure metrics, cholesterol, and lifestyle factors, with cholesterol and glucose levels particularly emphasized due to their poor replication in the synthetic dataset, as evidenced by the very low P-values and high error metrics (MSE, RMSE, MAE) recorded in Table 4.18.

Visual Insights and Correlation Analysis

The visual analysis of the distribution discrepancies for AE (Figure 4.54) and VAE (Figure 4.67) synthetic data highlights the extent of these variations, especially in features like alcohol and smoking for the VAE model, which exhibited the highest KS statistics, indicating major distributional differences. Similarly, the AE model showed the highest discrepancies in gender and active lifestyle indicators. Both models demonstrated exaggerated correlations between certain attributes within their datasets, such as gender and height, which suggest potential biases or overfitting in the data generation process.

Classification Performance and Predictive Accuracy

Both AE and VAE synthetic datasets generally offered improved classifier performance over the original dataset. Figure 4.62, Figure 4.64, and Figure 4.66 show that classifiers like Random Forest and XGB achieved higher AUC-ROC scores on synthetic datasets, indicating better predictive performance and a cleaner, more consistent representation of underlying disease patterns. This suggests that despite their discrepancies, synthetic datasets can potentially enhance machine learning applications by providing a more standardized data environment.

Our analysis confirms that both AE and VAE models are capable of generating useful synthetic datasets for cardiovascular disease studies, yet each has distinct areas requiring refinement. The AE model, detailed through statistical results in Table 4.16 and visual data in Figure 4.54 and Figure 4.55, needs improvements to better replicate lifestyle-related features and cholesterol levels accurately. Meanwhile, the VAE model, supported by data from Table 4.17 and Figure 4.67, and Figure 4.68, shows a need for better handling of blood pressure metrics and glucose levels. These findings underscore the need for ongoing enhancements in synthetic data generation techniques to ensure the datasets' reliability and applicability in healthcare research, particularly in maintaining an optimal balance between data utility and privacy. These efforts are crucial for ensuring that synthetic datasets can faithfully mimic real-world data, thus supporting robust and accurate predictive modeling in cardiovascular disease studies.

5.8 Overview of Findings from Privacy Risk Assessments on the Generated AE-Synthetic and VAE-Synthetic Cardiovascular Disease Data

5.8.1 Findings from Privacy Risk Assessments on AE Synthetic Cardiovascular Disease Data

Our study comprehensively evaluates the privacy preservation of the AE synthetic cardiovascular disease dataset, examining its resilience against various privacy risks through detailed risk assessments.

In singling-out risk assessment, both univariate and multivariate assessments, detailed in Tables 4.20 and Table 4.21, indicate extremely low privacy risks across different attack scenarios (1500 and 500 attacks), with success rates ranging from 0.0013 to 0.0038. These findings, visualized in Figures 4.79 and Figure 4.81, demonstrate the dataset's robustness in protecting against potential privacy breaches, even under rigorous testing conditions.

The linkability risk assessment in Table 4.22 shows minimal success rates for linkability attacks, even when the number of neighbors varies. With 10 neighbors, the main attack success rate was as low as 0.0021, dropping further to 0.0005 with 5 neighbors. The bar graph in Figure 4.83 and the pie charts in Figure 4.84 underline the dataset's strong anonymization capabilities, effectively preventing the linking of data points to individual identities.

We found that the inference risk assessment in Table 4.23 assesses the risks of inferring specific health attributes like height, weight, and aphi, where moderate vulnerabilities were noted. Conversely, attributes like age, gender, and cholesterol showed very low to no inference risk, affirming effective anonymization. The bar graph in Figure 4.85 and the pie charts in Figure 4.86 detail these risks and the success rates of various attack types, portraying a balanced profile of attack outcomes and highlighting the dataset's comprehensive security measures.

The extensive evaluations across multiple dimensions of privacy risks confirm the AE synthetic dataset's effectiveness in safeguarding privacy, with notable strengths in resisting both

singling-out and linkability attacks. While certain areas display vulnerabilities, particularly in inference risk for specific attributes, the implemented protective measures ensure that the dataset remains a reliable and secure tool for cardiovascular disease research. This robust analysis provides vital insights for stakeholders in healthcare analytics, emphasizing the strengths and areas for improvement in synthetic data generation technologies.

5.8.2 Findings from Privacy Risk Assessments on VAE Synthetic Cardiovascular Disease Data

Our analysis critically evaluates the privacy preservation of the VAE synthetic cardiovascular disease dataset through detailed risk assessments essential for its application in privacy-sensitive healthcare research.

The univariate and multivariate singling-out risk assessments in Tables 4.24 and Table 4.25 reveal extremely low privacy risks under various attack scenarios (1500 and 500 attacks), with all privacy risk assessments indicating zero risk. This underscores the dataset's robustness in preventing individual re-identification, further supported by the consistently low success rates illustrated in Figure 4.88. Linkability Risk Assessment, Table 4.26 shows minimal linkability risks, even as neighbor counts vary. Lower neighbor counts further decrease success rates, enhancing anonymization and demonstrating the dataset's effectiveness against linkability attacks (Figure 4.92). Table 4.27 assesses inference risks, identifying moderate vulnerabilities in attributes like height and blood pressure. Despite some risks, most attributes exhibit low to no inference risk, indicating strong privacy protections. Bar Graphs and Pie Charts (Figures 4.89, Figure 4.90, Figure 4.91, and Figure 4.93) visualize the success and failure rates of privacy attacks, consistently showing low success rates across various assessments. This visualization reinforces the synthetic data's capability to safeguard against privacy breaches effectively.

The comprehensive evaluations, documented in Tables 4.24 to Table 4.27 and visualized in Figures 4.88 to Figure 4.93, validate the VAE synthetic dataset's ability to protect privacy effectively. It excels particularly in resisting singling-out and linkability attacks, although some attributes show increased inference risk. Overall, the dataset remains a reliable tool for cardiovascular research within privacy-preserving frameworks, crucial for advancing healthcare analytics securely.

5.8.3 Comparative Analysis of Privacy Risk Assessment Across AE-Synthetic and VAE-Synthetic Cardiovascular Disease Data

Our project undertakes a critical evaluation of privacy preservation across two types of synthetic datasets for cardiovascular disease research: Autoencoder (AE)-Synthetic and Variational Autoencoder (VAE)-Synthetic. This comparison focuses on their respective capabilities to safeguard privacy through univariate, multivariate, linkability, and inference risk assessments, highlighting the strengths and potential vulnerabilities of each approach in maintaining data confidentiality.

Univariate Risk Assessment

- **AE-Synthetic Data:** The AE model demonstrated strong privacy preservation with extremely low risk of individual re-identification. Under both 1500 and 500 attack scenarios, the success rates were minimal, with main, baseline, and control attack success rates ranging between 0.0013 and 0.0038 (Tables 4.20 and Table 4.21, Figure 4.79).
- **VAE-Synthetic Data:** Similarly, the VAE model exhibited equally robust privacy measures. Regardless of the attack scenario, the privacy risks consistently registered at zero, indicating an optimal level of privacy preservation (Tables 4.24 and Table 4.25, Figure 4.88).

Multivariate Risk Assessment

- **AE-Synthetic Data:** Multivariate analyses maintained low success rates across various attack scenarios, confirming the data's integrity against complex privacy threats (Figure 4.81).
- **VAE-Synthetic Data:** The VAE model paralleled the AE model's performance in multivariate scenarios, showcasing negligible privacy risks and solidifying the data's robustness against sophisticated singling-out attempts (Figure 4.90).

Linkability Risk Assessment

- **AE-Synthetic Data:** The AE dataset showed formidable resistance to linkability attacks, especially as the number of neighbors was reduced, indicating strong anonymization techniques (Table 4.22, Figure 4.83 and Figure 4.84).
- **VAE-Synthetic Data:** The VAE dataset reported minimal linkability risks, similar to the AE dataset, with even lower success rates in scenarios with fewer neighbors, underscoring enhanced privacy protection (Figure 4.92).

Inference Risk Assessment

- **AE-Synthetic Data:** Certain attributes like height, weight, and arterial pressure showed moderate vulnerability, while others like age, gender, and cholesterol displayed very low risks (Table 4.23, Figure 4.85, and Figure 4.86).
- **VAE-Synthetic Data:** The VAE model revealed moderate vulnerabilities in attributes like height and blood pressure, indicating areas where privacy risks might be concentrated and needing cautious data handling (Table 4.27).

Visual Analyses and Further Insights: Both datasets employed visual tools effectively to illustrate the success and failure rates of various privacy attacks, providing intuitive insights into each dataset's defense mechanisms against potential privacy breaches.

Both AE and VAE synthetic datasets exhibit formidable privacy preservation capabilities, each excelling in resisting singling-out and linkability attacks, which is crucial for their use in sensitive healthcare research. However, both datasets also show some vulnerabilities in inference risks, necessitating ongoing enhancements to synthetic data generation technologies. These findings are vital for stakeholders in healthcare analytics, offering critical insights into the strengths and areas for improvement in maintaining data confidentiality in synthetic datasets.

5.9 Assessing the Fidelity and Utility of Autoencoder-Generated Synthetic Data in Mirroring the Statistical Characteristics of Original Lower Back Pain Datasets

The evaluation of AE synthetic lower back pain data focuses on its reliability compared to the original dataset, particularly underlining its application in handling chronic lower back pain—a prevalent condition requiring precise diagnostic models.

Statistical Analysis and Error Metrics Analysis: These revealed key differences in pelvic radius and degreespondylolisthesis with low KS-Test P-Values (<0.01), suggesting notable discrepancies that could impact clinical simulations as shown in Table 4.28. Regarding error metrics, high MSE, RMSE, and MAE, particularly for degreespondylolisthesis, point

to substantial deviations from the original data, impacting the clinical applicability of the synthetic dataset as detailed in Table 4.29.

Graphical Interpretations: We found that Figure 4.97 highlights significant differences in distributions and variances, such as in pelvic radius and degreespondylolisthesis, necessitating refinement in synthetic data generation. The scatter plot in Figure 4.98 shows alignment and variances across F-Test and T-Test results, emphasizing areas requiring cautious use of synthetic data. Moreover, detailed in Figure 4.99, higher error metrics for certain features underscore challenges in accurately replicating complex clinical data traits.

Correlation Matrix Analysis Insights: It is revealed from Figure 4.119 to Figure 4.101, that AE synthetic data typically shows stronger correlations than the original, possibly indicating overfitting or amplified underlying data patterns, which might affect its realism.

Classification and Predictive Analysis: Examination across various classifiers as documented in Figure 4.107 to Figure 4.115 shows that while some maintained or improved performance on synthetic data, others like the Decision Tree demonstrated substantial performance differences, highlighting potential data handling variations. Generally, higher AUC scores in synthetic data as seen in Figure 4.105 suggest smoother data that may simplify classification tasks but could also reduce the complexity needed for accurate real-world applications.

The thorough assessment of AE synthetic lower back pain data confirms its capabilities in mirroring general statistical properties and supporting classifier performance. Nonetheless, significant error metrics, distribution discrepancies, and occasionally exaggerated correlations point to crucial areas for improvement. This evaluation underscores the need for meticulous refinement of synthetic data techniques to enhance their clinical and research validity, ensuring they remain a reliable tool within privacy-preserving frameworks. The findings call for further research to optimize data generation processes, ensuring high fidelity and functional utility across various medical and research applications.

5.10 Overview of Findings from Privacy Risk Assessments on the Generated AE-Synthetic and VAE-Synthetic Lower Back Pain Data

The privacy preservation assessment of the 80% AE Synthetic Lower Back Pain data was conducted through univariate, multivariate, linkability, and inference risk evaluations. These assessments reveal the dataset's resilience against privacy breaches, highlighting both strengths and areas needing further protection.

Univariate Singling Out Risk Assessment: The univariate analysis, depicted in Table 4.32 and Figures 4.121 and Figure 4.122, showed that the dataset effectively resists privacy attacks under both 1500 and 500 attack scenarios, maintaining low success rates across attack types. Notably, the main attack's success rates slightly increased from 0.33% to 0.78% as attack volume decreased, indicating a potential vulnerability to more targeted attacks despite strong overall protections.

Multivariate Singling-Out Risk Assessment: Detailed in Table 4.33 and Figures 4.123 and Figure 4.124, the multivariate analysis exposed a moderate privacy risk that slightly increases with fewer attacks. The success rates for the main attack in these scenarios (19.54% for 1500 attacks and 21.62% for 500 attacks) suggest increased identification capabilities under concentrated attacks, pointing to potential vulnerabilities that could be exploited if not adequately addressed.

Linkability Risk Assessment: The linkability evaluations, summarized in Table 4.34 and

illustrated in Figures 4.125 and Figure 4.126, assessed the risk of correctly identifying individual records under settings of 10 and 5 neighbors. While the privacy risk remained low at 0.0 across both settings, the notable decrease in success rates with fewer neighbors highlights better privacy preservation under more stringent conditions.

Inference Risk Assessment: The inference risk analysis, shown in Table 4.35 and Figures 4.127 and Figure 4.129, identified the varied risk levels across different attributes of the dataset. Most attributes showed low to moderate risks; however, the sacrum angle exhibited a high risk (0.4064), marking it as a critical vulnerability that could be targeted for re-identification.

This comprehensive analysis underscores the AE synthetic lower back pain data's ability to manage privacy effectively across various dimensions. Despite robust protections, the escalation of risk under specific scenarios necessitates ongoing evaluations and enhancements to privacy safeguards, ensuring the dataset's safe use in privacy-sensitive applications.

5.11 Overview of Comparative Analysis of Classifiers Performance Accuracy on 80% AE and VAE Synthetic Datasets with Sourced Data at Kaggle Documented in Result Section

The evaluation of AE and VAE synthetic datasets against the original datasets in Table 4.40, Figure 4.11, Figure 4.34, Table 4.41, Figure 4.62, Figure 4.77, and Table 4.42, and Figure 4.106, using both AUC-ROC and 5-fold cross-validation metrics, has provided a comprehensive perspective on the performance of these models. These metrics results, commonly utilized in Kaggle competitions to benchmark model performance, highlight the capability of our models to distinguish between classes under various thresholds. The AUC-ROC metric is particularly crucial in clinical settings, where the sensitivity and specificity of diagnostic tools are paramount. For instance, similar to findings reported in Kaggle leaderboards, our models demonstrate robust discriminatory power, similar with slight variances that merit attention. You can view more details via Figure 4.25, Figure 4.26, Figure 4.24, and Figure 4.28

In contrast, the 5-fold cross-validation approach, another standard evaluation technique seen in Kaggle challenges, emphasized the consistency of model performance across different data subsets. This robustness is essential for verifying the generalizability of machine learning models in real-world applications. The cross-validation results revealed that while the AE synthetic datasets mimic the original data closely in many respects, there are notable discrepancies in performance metrics, suggesting areas for refinement in the synthetic generation process.

For example, divergences in AUC-ROC or cross-validation scores from those achieved on Kaggle could indicate nuances in the original data that the AE and VAE models have not fully captured. These discrepancies are particularly evident in complex disease datasets, where intricate data patterns require sophisticated modeling techniques to ensure accurate replication and prediction.

Implications for Practical Applications

Discussing these metrics within the context of their relevance to practical applications provides deeper insights into the suitability of synthetic data for operational use. Synthetic data holds tremendous potential in environments where data privacy is paramount, such as in healthcare. However, the slight differences in model performance, as indicated by our comparative analysis with Kaggle benchmarks, suggest a need for cautious integration into sensitive applications. The utilization of AUC-ROC and cross-validation metrics not only enriches our understanding of the comparative effectiveness of synthetic and original datasets but also guides future efforts to enhance the fidelity of synthetic data generation.

These findings underscore the need for ongoing optimization of synthetic data processes, ensuring that they provide robust, privacy-preserving, and accurate datasets for complex machine learning applications. As we continue to refine these models, insights drawn from benchmark platforms like Kaggle will remain invaluable in shaping the development of more sophisticated data synthesis tools.

5.12 Overview of Findings Between AE and VAE Models, and CTGAN and CopulaGAN Literature Review

As previously discussed in Chapters 2 and 4, by Maria Elinor Pedersen. [62] has demonstrated varying degrees of performance by CTGAN and CopulaGAN models across the same three different healthcare datasets as used in our study. This study extends these findings by employing the capabilities of AE and VAE models. The results underscore the superior strength of AE and VAE in generating high-quality synthetic data across multiple healthcare datasets, which is crucial for developing predictive models that are both effective and privacy-compliant. While Maria Elinor Pedersen's work with CTGAN and CopulaGAN provided foundational insights into the use of GANs for synthetic data generation, our findings suggest that AE and VAE offered enhanced performance, especially in scenarios requiring high data fidelity and privacy protection of the original data.

These findings suggest that classifiers trained on datasets generated by the AE and VAE models provide a more balanced approach to synthetic data generation, successfully mitigating the privacy concerns that were less adequately addressed by CTGAN and CopulaGAN. Such advancements are pivotal as they offer a pathway to deploying synthetic data in more sensitive applications, such as in clinical settings where data privacy is paramount.

The improved performance and privacy features of AE and VAE models call for further investigation into their deployment in real-world scenarios. Future research should explore the integration of these models into existing healthcare data systems to assess their practical utility and to refine their capabilities in line with evolving data protection regulations.

5.13 Future Work

The advancements we have achieved in enhancing the privacy preservation capabilities of our autoencoder and variational autoencoder models represent significant strides towards safer, more reliable synthetic data generation. However, the field of data privacy is dynamic, with new challenges continually emerging as technology and data use evolve. Recognizing this, there are several promising areas for future research and development that build upon our current work and push the boundaries of what is possible in privacy technology.

Exploration of Advanced Architectural Innovations: While we have made notable improvements to the architecture of our models, the potential of emerging neural network architectures offers a fertile ground for further exploration. Future work could include the integration of more sophisticated neural structures, such as transformer models or generative adversarial networks (GANs), which could offer new ways of enhancing privacy while possibly improving the utility of synthetic data.

Development of More Robust Privacy Metrics: Our use of the Anonymeter tool has been instrumental in assessing privacy risks. However, the development of more comprehensive privacy metrics that can provide deeper insights into potential vulnerabilities and the effectiveness of mitigation strategies is crucial. Future projects could focus on creating or refining tools that measure a broader spectrum of privacy aspects, including more nuanced

forms of data inference and re-identification risks.

Application to Diverse Data Types: Our current models have been primarily tested with specific types of data. Expanding this research to include a wider variety of data types, such as time-series, geographical, or even more complex structured data, could help in understanding the challenges and effectiveness of privacy-preserving techniques across different domains.

Real-World Implementation and Testing: Another significant area for future work involves the practical implementation of our models in real-world scenarios. This includes partnerships with industry to test the synthetic data in actual business processes or clinical settings, providing a clearer picture of their performance and utility outside of controlled experiments.

Continuous Learning and Adaptation: As data environments are not static, future iterations of our models could incorporate continuous learning mechanisms that adapt to changes in data patterns or privacy regulations. This would ensure that the synthetic data generation remains effective and compliant over time.

Ethical and Regulatory Considerations: Finally, as the field evolves, so too must our understanding of the ethical implications of synthetic data use. Future research should also include thorough evaluations of how synthetic data impacts individuals and communities, ensuring that our advances in privacy technology align with broader societal values and legal frameworks.

By pursuing these avenues, we can continue to refine our approaches, enhance the capabilities of our synthetic data models, and address the ever-changing landscape of privacy concerns. Our commitment to continuous improvement will not only advance the technical aspects of our work but also contribute to the development of more ethical and socially responsible data practices.

Chapter 6

Conclusion

Throughout this project, we embarked on an intricate exploration of synthetic data generation, focusing on the paramount importance of preserving privacy while maintaining high data utility. Employing Autoencoders (AE) and Variational Autoencoders (VAE), our efforts were directed towards not only understanding but also innovating at the intersection of data utility and privacy preservation within three critical healthcare datasets: obesity, cardiovascular disease, and lower back pain.

Our findings confirm that both AE and VAE models are exceptionally capable of generating synthetic datasets that mirror the statistical properties of original data, yet significantly enhance privacy. The rigorous evaluation through various statistical tests and privacy assessment tools, such as the Anonymeter, has substantiated the effectiveness of our model configurations. By meticulously adjusting the architecture of these models—particularly through modifications in the latent space dimensionality and the strategic use of Dense layers—we achieved a robust framework that adeptly balances the dual objectives of data utility and privacy.

The architectural refinements and iterative training adjustments were pivotal in our project. These modifications ensured that our models did not merely replicate data but did so with an acute awareness of privacy risks, effectively minimizing potential vulnerabilities. Our proactive approach, which included continual validation of privacy impact, allowed each iteration of the model development to move us closer to our goal of robust privacy protection. This dynamic process of assessment and adjustment highlighted our commitment to responsible data handling, aligning with stringent privacy requirements essential in sensitive applications.

Moreover, the comparative analysis with existing models like CTGAN and CopulaGAN emphasized the superior capabilities of AE and VAE in handling complex dataset nuances, particularly in ensuring privacy preservation without compromising data utility. This was evident from the performance metrics and classifier evaluations which consistently showed that synthetic data produced by our models maintained a high degree of similarity to real data, confirming their practical viability.

In conclusion, this project not only met its core objectives but also laid a strong foundation for future research in synthetic data generation. The insights gained and the methodologies developed provide a valuable framework for advancing synthetic data techniques that are both effective and secure. As we look forward, the lessons learned from this endeavor will undoubtedly influence ongoing and future efforts to refine and enhance privacy-preserving technologies in data-driven industries. Our journey through the complexities of data privacy and utility reaffirms our conviction in the potential of AE and VAE models to set new benchmarks in privacy-preserving synthetic data generation, heralding a new era of data security in healthcare and beyond.

Please visit our GitHub for the complete codebase and documentation:
https://github.com/Kelechiede/Enhancing_Data_Security_with_AE_and_VAE_Synthetic_Data_Generation

Bibliography

[1] Accountability Act. 'Health insurance portability and accountability act of 1996'. In: *Public law* 104 (1996), p. 191.

[2] M Arjovsky. 'S. Chintala i L. Bottou, Wasserstein GAN'. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. 2017, pp. 214–223.

[3] Martin Arjovsky. 'Soumith Chintala a Léon Bottou'. In: *Wasserstein GAN, arXiv* (2017).

[4] Dor Bank, Noam Koenigstein and Raja Giryes. 'Autoencoders'. In: *Machine learning for data science handbook: data mining and knowledge discovery handbook* (2023), pp. 353–374.

[5] Mrinal Kanti Baowaly et al. 'Synthesizing electronic health records using improved generative adversarial networks'. In: *Journal of the American Medical Informatics Association* 26.3 (2019), pp. 228–241.

[6] Siddharth Biswal et al. 'EVA: Generating longitudinal electronic health records using conditional variational autoencoders'. In: *Machine Learning for Healthcare Conference*. PMLR. 2021, pp. 260–282.

[7] Vladimir Bok and Jakub Langr. *GANs in Action: Deep learning with Generative Adversarial Networks*. Simon and Schuster, 2019.

[8] Stavroula Bourou et al. 'A review of tabular data synthesis using GANs on an IDS dataset'. In: *Information* 12.09 (2021), p. 375.

[9] Ramiro Camino, Christian Hammerschmidt and Radu State. 'Generating multi-categorical samples with generative adversarial networks'. In: *arXiv preprint arXiv:1807.01202* (2018).

[10] Gauthier Chassang. 'The impact of the EU general data protection regulation on scientific research'. In: *ecancermedicalscience* 11 (2017).

[11] Dingfan Chen et al. 'Gan-leaks: A taxonomy of membership inference attacks against generative models'. In: *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 2020, pp. 343–362.

[12] Richard J Chen et al. 'Synthetic data in machine learning for medicine and healthcare'. In: *Nature Biomedical Engineering* 5.6 (2021), pp. 493–497.

[13] Tianyue Cheng, Tianchi Fan and Landi Wang. 'Genetic Constrained Graph Variational Autoencoder for COVID-19 Drug Discovery'. In: *arXiv preprint arXiv:2104.11674* (2021).

[14] Edward Choi et al. 'Generating multi-label discrete patient records using generative adversarial networks'. In: *Machine learning for healthcare conference*. PMLR. 2017, pp. 286–305.

[15] Penny Chong et al. 'Simple and effective prevention of mode collapse in deep one-class classification'. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–9.

[16] João Coutinho-Almeida, Pedro Pereira Rodrigues and Ricardo João Cruz-Correia. 'GANs for tabular healthcare data generation: a review on utility and privacy'. In: *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*. Springer. 2021, pp. 282–291.

[17] Antonia Creswell et al. 'Generative adversarial networks: An overview'. In: *IEEE signal processing magazine* 35.1 (2018), pp. 53–65.

[18] Pádraig Cunningham, Matthieu Cord and Sarah Jane Delany. 'Supervised learning'. In: *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer, 2008, pp. 21–49.

[19] Ashok Cutkosky and Harsh Mehta. 'Momentum improves normalized sgd'. In: *International conference on machine learning*. PMLR. 2020, pp. 2260–2268.

[20] Sajad Darabi and Yotam Elor. 'Synthesising multi-modal minority samples for tabular data'. In: *arXiv preprint arXiv:2105.08204* (2021).

[21] Scott N Dean et al. 'PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction'. In: *Frontiers in microbiology* 12 (2021), p. 725727.

[22] Elise Devaux. *Presenting Anonymeter: The Tool for Assessing Privacy Risks in Synthetic Datasets*. https://www.anonos.com/blog/presenting-anonymeter-the-tool-for-assessing-privacy-risks-in-synthetic-datasets. Accessed: yyyy-mm-dd. 2023.

[23] Bin Ding, Huimin Qian and Jun Zhou. 'Activation functions and their characteristics in deep neural networks'. In: *2018 Chinese control and decision conference (CCDC)*. IEEE. 2018, pp. 1836–1841.

[24] Hongyuan Dong et al. 'Variational autoencoder for anti-cancer drug response prediction'. In: *arXiv preprint arXiv:2008.09763* (2020).

[25] John Duchi, Elad Hazan and Yoram Singer. 'Adaptive subgradient methods for online learning and stochastic optimization.' In: *Journal of machine learning research* 12.7 (2011).

[26] Cynthia Dwork. 'Differential privacy: A survey of results'. In: *International conference on theory and applications of models of computation*. Springer. 2008, pp. 1–19.

[27] Cristóbal Esteban, Stephanie L Hyland and Gunnar Rätsch. 'Real-valued (medical) time series generation with recurrent conditional gans'. In: *arXiv preprint arXiv:1706.02633* (2017).

[28] William Fedus et al. 'Many paths to equilibrium: GANs do not need to decrease a divergence at every step'. In: *arXiv preprint arXiv:1710.08446* (2017).

[29] Alvaro Figueira and Bruno Vaz. 'Survey on synthetic data generation, evaluation methods and GANs'. In: *Mathematics* 10.15 (2022), p. 2733.

[30] Clara García-Vicente et al. 'Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors'. In: *Applied Sciences* 13.7 (2023), p. 4119.

[31] Ghadeer Ghosheh, Jin Li and Tingting Zhu. 'A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation measures and data sources'. In: *arXiv preprint arXiv:2203.07018* (2022).

[32] Matteo Giomi et al. 'A unified framework for quantifying privacy risk in synthetic data'. In: *arXiv preprint arXiv:2211.10459* (2022).

[33] Ian Goodfellow et al. 'Generative adversarial nets'. In: *Advances in neural information processing systems* 27 (2014).

[34] Ian Goodfellow et al. 'Generative adversarial networks'. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[35] Mikel Hernadez et al. 'Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions'. In: *Methods of Information in Medicine* (2023).

[36] Mikel Hernandez et al. 'Synthetic data generation for tabular health records: A systematic review'. In: *Neurocomputing* 493 (2022), pp. 28–45.

[37] Yufei Huang and Jianqiu Zhang. 'Exploring factor structures using variational autoencoder in personality research'. In: *Frontiers in psychology* 13 (2022), p. 863926.

[38] Sergey Ioffe and Christian Szegedy. 'Batch normalization: Accelerating deep network training by reducing internal covariate shift'. In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[39] Leonid Joffe. 'Transfer Learning for Tabular Data'. In: *Authorea Preprints* (2023).

[40] James Jordon, Jinsung Yoon and Mihaela Van Der Schaar. 'PATE-GAN: Generating synthetic data with differential privacy guarantees'. In: *International conference on learning representations*. 2018.

[41] Diederik P Kingma and Jimmy Ba. 'Adam: A method for stochastic optimization'. In: *arXiv preprint arXiv:1412.6980* (2014).

[42] Diederik P Kingma, Max Welling et al. 'An introduction to variational autoencoders'. In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.

[43] Will Koehrsen. 'Overfitting vs. underfitting: A complete example'. In: *Towards Data Science* 405 (2018).

[44] Anders Krogh. 'What are artificial neural networks?' In: *Nature biotechnology* 26.2 (2008), pp. 195–197.

[45] Karol Kurach et al. 'A large-scale study on regularization and normalization in GANs'. In: *International conference on machine learning*. PMLR. 2019, pp. 3581–3590.

[46] Jakub Langr and Vladimir Bok. 'GANs in action: deep learning with generative adversarial networks'. In: *(No Title)* (2019).

[47] Gael Lederrey, Tim Hillel and Michel Bierlaire. 'DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data'. In: *arXiv preprint arXiv:2203.03489* (2022).

[48] Feng Li et al. 'Input layer regularization of multilayer feedforward neural networks'. In: *IEEE Access* 5 (2017), pp. 10979–10985.

[49] Hongming Li, Shujian Yu and Jose Principe. 'Causal recurrent variational autoencoder for medical time series generation'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 7. 2023, pp. 8562–8570.

[50] Zinan Lin, Vyas Sekar and Giulia Fanti. 'Why spectral normalization stabilizes gans: Analysis and improvements'. In: *Advances in neural information processing systems* 34 (2021), pp. 9625–9638.

[51] Claire Little et al. 'Generative adversarial networks for synthetic data generation: a comparative study'. In: *arXiv preprint arXiv:2112.01925* (2021).

[52] Chaoyue Liu and Mikhail Belkin. 'Accelerating sgd with momentum for over-parameterized learning'. In: *arXiv preprint arXiv:1810.13395* (2018).

[53] Xuan Liu et al. 'Medical image compression based on variational autoencoder'. In: *Mathematical Problems in Engineering* 2022 (2022).

[54] Persevearance Marecha and Lu Ye. 'Generation and Evaluation of Tabular Data in Different Domains Using Gans'. In: *Asian Journal of Research in Computer Science* 16.1 (2023), pp. 15–27.

[55] Umberto Michelucci. 'An introduction to autoencoders'. In: *arXiv preprint arXiv:2201.03898* (2022).

[56] Mehdi Mirza and Simon Osindero. 'Conditional generative adversarial nets'. In: *arXiv preprint arXiv:1411.1784* (2014).

[57] Giannis Nikolentzos et al. 'Synthetic electronic health records generated with variational graph autoencoders'. In: *npj Digital Medicine* 6.1 (2023), p. 83.

[58] Marmar Orooji, Seyedeh Shaghayegh Rabbanian and Gerald M Knapp. 'Flexible adversary disclosure risk measure for identity and attribute disclosure attacks'. In: *International Journal of Information Security* 22.3 (2023), pp. 631–645.

[59] Noseong Park et al. 'Data synthesis based on generative adversarial networks'. In: *arXiv preprint arXiv:1806.03384* (2018).

[60] Neha Patki, Roy Wedge and Kalyan Veeramachaneni. 'The synthetic data vault'. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2016, pp. 399–410.

[61] Karl Pearson. 'On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling'. In: *Philosophical Magazine Series 5* 50.302 (1900), pp. 157–175.

[62] Maria Elinor Pedersen. 'Exploring the Value of GANs for Synthetic Tabular Data Generation in Healthcare with a Focus on Data Quality, Augmentation, and Privacy'. MA thesis. Oslomet-storbyuniversitetet, 2023.

[63] Ekachai Phaisangittisagul. 'An analysis of the regularization between L2 and dropout in single hidden layer neural network'. In: *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*. IEEE. 2016, pp. 174–179.

[64] Daniil Polykovskiy et al. 'Entangled conditional adversarial autoencoder for de novo drug discovery'. In: *Molecular pharmaceutics* 15.10 (2018), pp. 4398–4405.

[65] David Pratella et al. 'A survey of autoencoder algorithms to pave the diagnosis of rare diseases'. In: *International journal of molecular sciences* 22.19 (2021), p. 10891.

[66] Alec Radford, Luke Metz and Soumith Chintala. 'Unsupervised representation learning with deep convolutional generative adversarial networks'. In: *arXiv preprint arXiv:1511.06434* (2015).

[67] Amirarsalan Rajabi and Ozlem Ozmen Garibay. 'Tabfairgan: Fair tabular data generation with generative adversarial networks'. In: *Machine Learning and Knowledge Extraction* 4.2 (2022), pp. 488–501.

[68] Sundaramoorthy Rajasekaran and GA Vijayalakshmi Pai. *Neural networks, fuzzy systems and evolutionary algorithms: Synthesis and applications*. PHI Learning Pvt. Ltd., 2017.

[69] Raúl Rojas. *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[70] Sebastian Ruder. 'An overview of gradient descent optimization algorithms'. In: *arXiv preprint arXiv:1609.04747* (2016).

[71] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams et al. *Learning internal representations by error propagation*. 1985.

[72] Maziar Sanjabi et al. 'On the convergence and robustness of training gans with regularized optimal transport'. In: *Advances in Neural Information Processing Systems* 31 (2018).

[73] Akash Srivastava et al. 'Veegan: Reducing mode collapse in gans using implicit variational learning'. In: *Advances in neural information processing systems* 30 (2017).

[74] Theresa Stadler, Bristena Oprisanu and Carmela Troncoso. 'Synthetic Data – Anonymisation Groundhog Day'. In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 1451–1468. ISBN: 978-1-939133-31-1. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

[75] Jonas Teuwen and Nikita Moriakov. 'Convolutional neural networks'. In: *Handbook of medical image computing and computer assisted intervention*. Elsevier, 2020, pp. 481–501.

[76] Tijmen Tieleman. 'Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude'. In: *COURSERA: Neural networks for machine learning* 4.2 (2012), p. 26.

[77] Shivani Tomar and Ankit Gupta. 'A Review on Mode Collapse Reducing GANs with GAN's Algorithm and Theory'. In: *GANs for Data Augmentation in Healthcare* (2023), pp. 21–40.

[78] Muhammad Uzair and Noreen Jamil. 'Effects of hidden layers on the efficiency of neural networks'. In: *2020 IEEE 23rd international multitopic conference (INMIC)*. IEEE. 2020, pp. 1–6.

[79] Lu Wang, Wei Zhang and Xiaofeng He. 'Continuous patient-centric sequence generation via sequentially coupled adversarial learning'. In: *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part II 24*. Springer. 2019, pp. 36–52.

[80] Yang Wang. 'A mathematical introduction to generative adversarial nets (GAN)'. In: *arXiv preprint arXiv:2009.00169* (2020).

[81] Jinhong Wu et al. 'Interpretation for variational autoencoder used to generate financial synthetic tabular data'. In: *Algorithms* 16.2 (2023), p. 121.

[82] Lei Xu and Kalyan Veeramachaneni. 'Synthesizing tabular data using generative adversarial networks'. In: *arXiv preprint arXiv:1811.11264* (2018).

[83] Lei Xu et al. 'Modeling tabular data using conditional gan'. In: *Advances in neural information processing systems* 32 (2019).

[84] Andrew Yale et al. 'Generation and evaluation of privacy preserving synthetic health data'. In: *Neurocomputing* 416 (2020), pp. 244–255.

[85] Jinsung Yoon, Daniel Jarrett and Mihaela Van der Schaar. 'Time-series generative adversarial networks'. In: *Advances in neural information processing systems* 32 (2019).

[86] Yuheng Zhang et al. 'The secret revealer: Generative model-inversion attacks against deep neural networks'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 253–261.

[87] Zhengli Zhao et al. 'Image augmentations for gan training'. In: *arXiv preprint arXiv:2006.02595* (2020).

[88] Zilong Zhao et al. 'Ctab-gan: Effective table data synthesizing'. In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 97–112.

Appendices

Explanation of Calculation Process for Presentations of All Results by Pie Charts

The calculation processes outlined in this section are consistent across the pie chart representations for three different datasets: obesity, cardiovascular disease, and lower back pain. To avoid redundancy, we demonstrate the processes using the VAE-Synthetic Lower Back Pain dataset as an example.

Explanation of Calculation Process for Pie Charts Representing Success and Failure Rates

Given the success rates for the Main, Baseline, and Control attacks on the VAE Synthetic Lower Back Pain dataset, we calculate the failure rates by subtracting each success rate from 100%.

Main Attack

$$\begin{aligned} \text{Success Rate} &= 74.3\% \\ \text{Failure Rate} &= 100\% - 74.3\% = 25.7\% \end{aligned}$$

Baseline Attack

$$\begin{aligned} \text{Success Rate} &= 45.4\% \\ \text{Failure Rate} &= 100\% - 45.4\% = 54.6\% \end{aligned}$$

Control Attack

$$\begin{aligned} \text{Success Rate} &= 66.7\% \\ \text{Failure Rate} &= 100\% - 66.7\% = 33.3\% \end{aligned}$$

Explanation of Calculation Process for Pie Charts Representing Overall Success versus Failure Rates

The overall success rate is calculated by averaging the success rates of the Main, Baseline, and Control attacks. The overall failure rate is then determined by subtracting the overall success rate from 100%.

$$\begin{aligned}\text{Overall Success Rate} &= \frac{\text{Main Attack Success} + \text{Baseline Attack Success} + \text{Control Attack Success}}{3} \\ &= \frac{74.3\% + 45.4\% + 66.7\%}{3} \\ &= \frac{186.4\%}{3} \\ &= 62.13\%\end{aligned}$$

$$\begin{aligned}\text{Overall Failure Rate} &= 100\% - \text{Overall Success Rate} \\ &= 100\% - 62.13\% \\ &= 37.87\%\end{aligned}$$

Mathematical Formulation: The overall success rate is the average of the three provided success rates. The overall failure rate is computed by subtracting the overall success rate from 100%.
Detailed Calculation Steps: Each step in the calculation is broken down, providing a clear pathway from individual rates to overall rates.
Result Interpretation: These calculations help illustrate the dataset's overall security by showing the percentage of attacks that do not result in success, thus demonstrating the effectiveness of the dataset's privacy safeguards.
This section should be placed in the appendix of your LaTeX document, ensuring it is properly referenced from the results section where the pie charts are discussed. You can link to this appendix section from the results chapter by referring to it using the '6' command in your discussion, guiding readers to the detailed mathematical underpinnings of the analyses presented in the pie charts.

Classification Report Metrics Detailed Analysis

Precision

Precision measures the accuracy of positive predictions. It is defined as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Positives (FP)}} \tag{6.1}$$

Recall (Sensitivity)

Recall measures the model's ability to find all the relevant cases. It is defined as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Negatives (FN)}} \tag{6.2}$$

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between them:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.3}$$

Accuracy

Accuracy measures the overall correctness of the model:

$$\text{Accuracy} = \frac{\text{True Positives (TP) + True Negatives (TN)}}{\text{Total Observations}} \tag{6.4}$$