RESEARCH ARTICLE

VOL. 46 | NO. 1 | **JANUARY-APRIL 2025** | PP 1-28

Revista Mexicana de Ingeniería Biomédica

https://dx.doi.org/10.17488/RMIB.46.1.3

E-LOCATION ID: e1480

Synthetic Data Generation for Pediatric Diabetes Research Using GANs and WGANs

Generación de Datos Sintéticos para la Investigación de la Diabetes Infantil Usando GANs y WGANs

Antonio García-Domínguez[1], *Carlos E. Galván-Tejada*[1], *Rafael Magallanes-Quintanar*[1] ✉, *Miguel Cruz-López*[2], *Miguel Alexander Vázquez-Moreno*[2], *Erika Acosta-Cruz*[3]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas - México

[2] Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Ciudad de México - México

[3] Universidad Autónoma de Coahuila, Departamento de Biotecnología, Coahuila - México

ABSTRACT

Pediatric diabetes research is often constrained by data scarcity, hindering the development of accurate predictive models for clinical applications. This study addresses this limitation by evaluating the effectiveness of Generative Adversarial Networks (GANs) and Wasserstein GANs (WGANs) in generating synthetic datasets that replicate the statistical properties of real pediatric diabetes data. A structured methodology was applied, incorporating preprocessing, model design, and dual evaluation metrics: Jensen-Shannon and Kullback-Leibler divergences for statistical fidelity, and a classification model to assess practical utility. Results demonstrate that both models produce high-fidelity synthetic datasets, with WGANs showing superior performance in capturing complex patterns due to improved training stability. Nonetheless, challenges remain in replicating the inherent variability of pediatric data, influenced by growth and developmental factors. This work highlights the potential of synthetic data to augment pediatric diabetes datasets, facilitating the development of robust and generalizable predictive models. Limitations include the dependency on initial data quality and the specificity of the models to pediatric datasets. By addressing critical gaps in data availability, this study contributes to advancing AI-driven healthcare solutions in pediatric diabetes research.

KEYWORDS: Generative Adversarial Networks (GANs), pediatric diabetes, synthetic data generation, Wasserstein GANs

RESUMEN

La investigación en diabetes pediátrica a menudo está limitada por la escasez de datos, lo que dificulta el desarrollo de modelos predictivos precisos para aplicaciones clínicas. Este estudio aborda esta limitación evaluando la efectividad de las Redes Generativas Antagónicas (GANs) y las Wasserstein GANs (WGANs) para generar conjuntos de datos sintéticos que replican las propiedades estadísticas de los datos reales de diabetes pediátrica. Se aplicó una metodología estructurada que incluye el preprocesamiento, diseño de modelos y métricas de evaluación dual: divergencias de Jensen-Shannon y Kullback-Leibler para evaluar la fidelidad estadística, y un modelo de clasificación para evaluar la utilidad práctica. Los resultados demuestran que ambos modelos generan datos sintéticos de alta fidelidad, siendo las WGANs superiores en la captura de patrones complejos gracias a su estabilidad de entrenamiento mejorada. Sin embargo, persisten desafíos para replicar la variabilidad inherente de los datos pediátricos, influida por el crecimiento y los factores de desarrollo. Este trabajo resalta el potencial de los datos sintéticos para aumentar los conjuntos de datos de diabetes pediátrica, facilitando el desarrollo de modelos predictivos robustos y generalizables. Las limitaciones incluyen la dependencia de la calidad de los datos iniciales y la especificidad de los modelos a los conjuntos de datos pediátricos. Este estudio contribuye a cerrar brechas críticas en la disponibilidad de datos, impulsando soluciones de salud personalizadas basadas en inteligencia artificial para la investigación en diabetes pediátrica.

PALABRAS CLAVE: diabetes infantil, generación de datos sintéticos, Redes Generativas Adversarias (GANs), Redes Generativas Adversarias de Wasserstein (WGANs)

Corresponding author

TO: Rafael Magallanes-Quintanar

INSTITUTION:UNIVERSIDAD AUTÓNOMA DE ZACATECAS

ADDRESS: JARDÍN JUAREZ 147, CENTRO, ZACATECAS 98000, ZAC, MÉXICO

EMAIL: tiquis@uaz.edu.mx

Received:

15 October 2024

Accepted:

26 December 2024

Published:

07 February 2025

INTRODUCTION

In recent years, artificial intelligence (AI) has achieved considerable advancements in the healthcare sector, introducing groundbreaking methods for disease diagnosis, prediction, and management. AI-driven models, particularly those utilizing machine learning (ML), have become indispensable tools in enhancing the accuracy of medical diagnoses and in personalizing treatment plans[1][2][3]. These advancements are especially crucial in the management of chronic diseases such as diabetes, where early detection and precise management are vital to improving patient outcomes and quality of life[4][5]. The success of AI in healthcare, however, is heavily dependent on the availability of large, high-quality datasets, which are often difficult to obtain due to various constraints including privacy concerns and the logistical challenges of data collection[6].

Diabetes, a chronic metabolic disorder, requires diligent management and has become a significant global health crisis. It is one of the leading causes of death and disability worldwide, affecting over 537 million adults and contributing to millions of deaths each year[7]. The burden of diabetes is escalating rapidly, with the prevalence expected to rise substantially in the coming decades. The disease is a major contributor to heart disease, stroke, kidney failure, blindness, and lower limb amputations, imposing enormous strain on healthcare systems and economies globally[8]. Without proper management, diabetes can lead to severe and life-threatening complications, making it a critical focus of public health efforts worldwide. Within this global context, diabetes in specific populations, such as children, demands particular attention due to the unique challenges it presents.

Pediatric diabetes, encompassing both type 1 and type 2 diabetes, presents distinct challenges due to its early onset and the requirement for lifelong care. Type 1 diabetes, the most common form in children, is an autoimmune condition that results in the destruction of insulin-producing beta cells in the pancreas, needing continuous insulin therapy for survival[9]. The increasing incidence of pediatric diabetes, particularly type 1 diabetes, is a growing concern, as it often manifests during critical developmental periods, heightening the risk of long-term complications such as cardiovascular disease, neuropathy, and retinopathy[10]. Additionally, the rise of type 2 diabetes in children, largely driven by the global obesity epidemic, adds further complexity to disease management as it combines insulin resistance with the challenges of ongoing growth and development[11].

Effective management and early intervention in pediatric diabetes hinge on the development of accurate and reliable predictive models. These models are essential for identifying at-risk individuals, optimizing treatment strategies, and ultimately improving patient outcomes. However, the scarcity of large, high-quality datasets in pediatric diabetes research presents a significant barrier to the creation of such models. The challenges of collecting comprehensive data in children are compounded by ethical concerns, including the need to protect patient privacy and obtain informed consent, as well as the natural variability in children's growth patterns, which introduces additional complexity into data collection and analysis. As a result, researchers often must work with small, inconsistent datasets that are not sufficient to train robust machine learning models.

To address these challenges, synthetic data generation has emerged as a promising solution. Synthetic data are artificially created datasets that replicate the statistical properties of real-world data, allowing researchers to augment existing datasets or generate new ones when real data are insufficient or difficult to obtain[12][13]. This approach is particularly valuable in pediatric diabetes research, where data limitations can significantly hinder the development of effective AI models. By generating synthetic data, researchers can overcome the constraints imposed by small and inconsistent datasets, thereby enhancing the robustness and generalizability of predictive models. This not only eases the development of more accurate models but also helps in exploring a wider range of scenarios and outcomes that may be underrepresented in real-world data.

Among the various techniques for generating synthetic data, Generative Adversarial Networks (GANs) and their variant, Wasserstein GANs (WGANs), have gained prominence due to their ability to produce realistic and high-quality synthetic data[10]. GANs consist of two neural networks, a generator and a discriminator, which work in tandem to create data that closely mimic real-world observations. The generator is tasked with producing synthetic data, while the discriminator evaluates the realism of the generated data, driving the generator to continuously improve its outputs[14]. WGANs, an extension of traditional GANs, address some of the limitations of GANs by employing the Wasserstein distance as a measure of how well the generated data match the real data. This approach leads to a more stable and effective training process, making WGANs particularly suitable for generating synthetic datasets in complex and high-dimensional domains like medical data[15].

In the context of pediatric diabetes, the application of GANs and WGANs for generating synthetic data is particularly promising. These models are adept at capturing the intricate and multi-dimensional relationships inherent in medical data, making them ideal for creating synthetic datasets that can be used to train AI models. However, generating synthetic data is only one part of the equation, it is equally important to ensure that the synthetic data can serve as a valid substitute for real-world data. The Jensen-Shannon divergence[16] and the Kullback-Leibler divergence [17] evaluate the similarity between the distributions of synthetic and real data[18]. The Jensen-Shannon divergence, due to its symmetric and bounded nature, serves as a general measure of how closely the synthetic data align with the real data, with lower values indicating a closer match. The Kullback-Leibler divergence offers additional insight by highlighting specific areas where the synthetic data may differ from the real data. Additionally, the synthetic data's utility is assessed through a classification model trained on both real and synthetic datasets, using metrics such as accuracy, precision, recall, F1-score, and ROC AUC. This combined approach ensures a comprehensive evaluation of the synthetic data, both in terms of its statistical alignment with real data and its practical applicability in AI model training and medical research.

Building on the importance of evaluating synthetic data with robust metrics, next-generation sequencing (NGS) has significantly contributed to the field of pediatric diabetes by enabling the identification of genetic variants associated with the disease. NGS has been used to uncover key insights into monogenic diabetes and gene-environment interactions, providing a deeper understanding of disease mechanisms and potential therapeutic targets [19]. However, the vast and heterogeneous datasets generated by NGS pose challenges for integration into predictive models, largely due to insufficient annotation and the complexity of multi-omics data[20]. Synthetic data generation, as proposed in this study, offers a complementary approach to address these limitations by augmenting NGS-derived datasets and enhancing their utility for machine learning applications. By employing GANs and WGANs, the synthetic datasets created in this work aim to align closely with real-world data distributions while improving their applicability for predictive modeling in pediatric diabetes research. These challenges underscore the critical need for innovative approaches like synthetic data generation, which not only augment NGS-derived datasets but also enable their integration into machine learning models for predictive analysis.

The importance of generating high-quality synthetic data has been underscored in recent studies, which emphasize the need for advanced synthetic data generation techniques to overcome the inherent limitations of real-world data collection[21][22]. For instance, the application of GANs in generating synthetic diabetes data has been successfully demonstrated in various studies, including work by García-Domínguez *et al.*[4]. This study addresses the critical issue of data scarcity in pediatric diabetes by generating synthetic datasets using both GAN and WGAN methodologies. The quality of the generated data is rigorously evaluated using both the Jensen-Shannon divergence and the Kullback-Leibler divergence to ensure close alignment with the characteristics of real-world data. The aim is to provide valuable insights into effective practices for synthetic data generation in medical research, ultimately supporting the development of AI-driven healthcare solutions that enhance the management and outcomes of pediatric diabetes.

The main contribution of this work lies in the evaluation of GANs and WGANs for generating synthetic data specifically for pediatric diabetes, a domain that remains underexplored despite its clinical importance. Unlike previous studies that focus solely on statistical metrics, this study introduces a dual evaluation approach, combining traditional statistical divergence measures (Jensen-Shannon and Kullback-Leibler) with a supervised classification model. This comprehensive validation not only ensures the fidelity of the generated data but also demonstrates its practical applicability for predictive tasks, addressing a critical gap in the generation and evaluation of synthetic pediatric datasets.

This paper is organized as follows: The “Materials and Methods” section describes the pediatric diabetes dataset and the preprocessing steps for synthetic data generation. It then outlines the architectures of the GAN and WGAN models and the techniques employed in their design. Additionally, the section details the methods used to evaluate the quality of the generated data, including both distributional similarity metrics (Jensen-Shannon and Kullback-Leibler divergences) and the performance of a classification model trained on real and synthetic data. The “Results and Discussion” section presents the outcomes of the synthetic data generation process, including a comparison of the two methods based on these evaluation approaches. Finally, the “Conclusions” section summarizes the key findings, discusses implications for future research, and highlights the potential of synthetic data in enhancing AI-driven healthcare solutions.

MATERIALS AND METHODS

This section presents the foundational methods and concepts employed in this study to generate and evaluate synthetic data for pediatric diabetes research. It begins with a comprehensive description of the dataset, highlighting the key clinical, biochemical, and genetic variables that are crucial for understanding diabetes in children. This is followed by an introduction to Generative Adversarial Networks (GANs) and their variant, Wasserstein GANs (WGANs), outlining their theoretical basis, mechanisms, and advantages in generating realistic synthetic data that mirrors complex medical datasets. Additionally, the section provides an overview of the evaluation metrics, including the Jensen-Shannon and Kullback-Leibler divergences, which are critical for assessing the quality and fidelity of the synthetic data relative to the original data, as well as the performance metrics derived from a classification model, which evaluate the practical applicability of the synthetic data in AI-driven tasks.

Figure 1 illustrates the methodological framework employed in this study. The process begins with the input of the original pediatric diabetes dataset, which undergoes preprocessing steps such as feature filtering, handling data incompleteness, and normalization to ensure the quality and consistency of the data. Subsequently, synthetic data is generated using two models: Generative Adversarial Networks (GAN) and Wasserstein GANs (WGAN). Finally, the synthetic datasets are evaluated through two complementary approaches: statistical divergence metrics (Jensen-Shannon and Kullback-Leibler) and a supervised classification model (Random Forest), which assesses the practical utility of the generated data. This pipeline provides a structured and comprehensive methodology for generating high-quality synthetic data tailored to the complexities of pediatric diabetes.

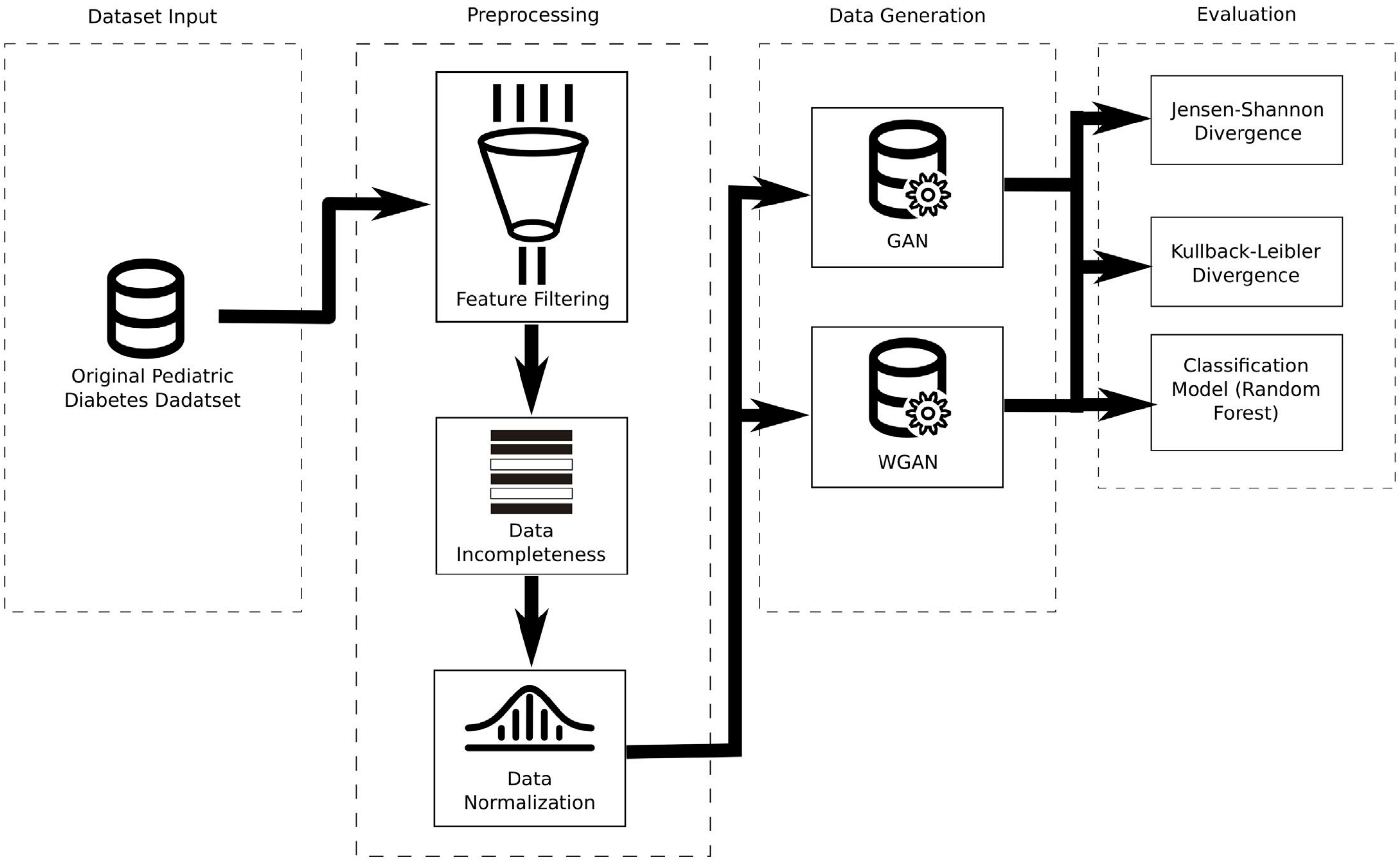

Dataset Input
Preprocessing
Data Generation
Evaluation
Original Pediatric Diabetes Dadatset
Feature Filtering
Data Incompleteness
Data Normalization
GAN
WGAN
Jensen-Shannon Divergence
Kullback-Leibler Divergence
Classification Model (Random Forest)

FIGURE 1. **Methodological framework for synthetic data generation and evaluation, including dataset input, preprocessing, data generation (GAN and WGAN), and evaluation through divergence metrics and classification modeling.**

Original dataset description

The dataset used in this study consists of 834 records related to pediatric diabetes patients. It includes a range of clinical, biochemical, and genetic variables that are crucial for understanding various aspects of diabetes in children. The dataset encompasses 22 features, each representing a different attribute relevant to the diagnosis and management of diabetes. These features are summarized in Table 1.

TABLE 1. **Overview of Features in the Pediatric Diabetes Dataset.**

Variable	Description
Location	Geographic location of the patient or the clinical study site.
StudyProject	The project or study identifier.
RecordID	A unique identifier for each patient record.
Sex	A binary variable indicating the sex of the patient, where '0' represents females (girls) and '1' represents males (boys).
Age	The age of the patient, measured in years.
BMI	Body Mass Index (BMI), which is a measure of body fat based on height and weight.
BMI_Z_Score	The z-score for BMI, which provides a standardized measure of BMI relative to a reference population.
BodyWeightStatus	A categorical variable indicating body weight classification, where '0' represents normal weight and '1' represents obesity.
SystolicBP	Systolic blood pressure, measured in millimeters of mercury (mmHg).
DiastolicBP	Diastolic blood pressure, also measured in millimeters of mercury (mmHg).
Glucose	Blood glucose level, measured in milligrams per deciliter (mg/dL).
TotalCholesterol	Total cholesterol level, measured in milligrams per deciliter (mg/dL).
Triglycerides	Triglyceride level, measured in milligrams per deciliter (mg/dL).
HDL_Cholesterol	High-density lipoprotein cholesterol (HDL-C) level, measured in milligrams per deciliter (mg/dL).
LDL_Cholesterol	Low-density lipoprotein cholesterol (LDL-C) level, measured in milligrams per deciliter (mg/dL).
Insulin	Insulin level, measured in micrograms per milliliter (g/mL).
HOMA_IR	Homeostatic Model Assessment for Insulin Resistance (HOMA-IR), which is used to evaluate insulin resistance.
SalivaryAmylase	Salivary amylase activity, measured in units per liter (UI/L).
PancreaticAmylase	Pancreatic amylase activity, measured in units per liter (UI/L).
TotalAmylase	Total amylase activity, measured in units per liter (UI/L).
CNVs_SalivaryAmylase	Copy number variations of the gene associated with salivary amylase production
CNVs_PancreaticAmylase	Copy number variations of the gene associated with pancreatic amylase production

Data corresponds to Mexican patients and were collected at the General Hospital *"Centro Médico Siglo XXI"* of the *Instituto Mexicano del Seguro Social (IMSS).* All participants provided informed consent prior to their inclusion in the study, titled *"Variación y funcionalidad del número de copias del gen de amilasa (AMY1, AMY2) y su asociación con microbiota intestinal en obesidad infantil"*. The study adhered to the principles outlined in the Declaration of Helsinki, and the protocol received approval from the Ethics Committee of the *"Instituto Mexicano del Seguro Social"* and the *"Comisión Nacional de Investigación Científica"* (R-2016-785-097).

Data preprocessing

Data preprocessing is a crucial step in preparing the dataset for synthetic data generation, ensuring that the models employed can effectively learn and generate high-quality synthetic samples that closely resemble real-world data. To achieve this, it is necessary to implement a series of systematic preprocessing steps, including the removal of irrelevant variables, the handling of data incompleteness, and the normalization of continuous variables. These procedures are outlined in detail below.

Feature filtering

In this study, it is necessary to filter the dataset to eliminate features that are not relevant to the objectives. The features 'Location', 'StudyProject', and 'RecordID' are removed, as they do not provide meaningful information for the analysis of pediatric diabetes. 'Location' and 'StudyProject' are administrative variables without clinical, biochemical, or genetic relevance, while 'RecordID' serves merely as a unique identifier with no analytical value. Removing these features reduces dimensionality, thereby simplifying the dataset and enhancing both the efficiency and performance of the models. In the context of machine learning and data science, reducing dimensionality is a critical step that improves model interpretability and mitigates the risk of overfitting by eliminating noise and irrelevant data[23][24].

Data incompleteness

Addressing missing values is essential in preparing medical datasets, where incomplete data can significantly impact the performance and reliability of machine learning models[25]. In this study, data incompleteness is managed through mean imputation, a technique widely employed for handling missing data, particularly when dealing with continuous variables. This approach involves replacing missing values with the mean of the corresponding feature, thereby preserving the overall central tendency and distribution of the data[26].

Mean imputation is applied by calculating the mean of each feature with missing values and substituting this mean for the missing entries[27]. Mathematically, for a given feature X_j with missing values, the imputed value $\hat{x_{ij}}$ for any missing observation x_{ij} is given by Equation (1).

$$\hat{x}_{ij} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{1}$$

where n represents the total number of observations for which the feature X_j is not missing. This method assumes that the missing data are Missing At Random (MAR), implying that the probability of missingness is not related to the missing values themselves but may be related to the observed data. Mean imputation preserves the mean of the observed data and minimizes the impact on the overall distributional properties, maintaining the variance structure to some extent. However, it may reduce the natural variability of the dataset by introducing repeated values, which is an acceptable trade-off in this context due to the small proportion of missing data[28][29][30].

While other imputation methods, such as median or mode imputation, can handle non-normally distributed data or categorical variables, mean imputation is chosen for its computational efficiency and its ability to maintain the central tendency of the dataset. This ensures that the imputations do not introduce significant bias, allowing the synthetic data generation models to learn from a complete and consistent dataset while preserving its original characteristics as closely as possible.

Data normalization

Data normalization is a fundamental step in preparing datasets for machine learning models, particularly neural networks, which are highly sensitive to the scale of input data[31]. In this study, Min-Max Scaling is employed to normalize the continuous variables, transforming each feature to a fixed range, typically [0, 1]. This normalization technique is defined by Equation (2).

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

where X represents the original feature values, X_{min} is the minimum value, and X_{max} is the maximum value of the feature[32]. Min-Max Scaling is chosen because it preserves the original distribution shape while ensuring that all features contribute equally to the model's learning process. This method is particularly advantageous over other normalization techniques, such as Z-score normalization, which centers the data around the mean and scales based on standard deviation but can be more sensitive to outliers and does not always maintain the relative spacing of values[33][34].

The use of Min-Max Scaling is especially beneficial in the context of Generative Adversarial Networks (GANs) and Wasserstein GANs (WGANs). These models require a stable training process to effectively generate high-quality synthetic data, and normalization helps by ensuring that gradients do not vanish or explode due to disparate feature scales. By standardizing the input data, the models can learn underlying patterns more effectively, thereby improving the robustness and quality of the synthetic data generated[35].

Generative models

The scarcity and variability of high-quality pediatric diabetes data present significant challenges in developing reliable predictive models. To mitigate these issues, various generative models have been developed to create synthetic datasets that replicate the statistical properties and complex relationships found in real-world data. In this study, Generative Adversarial Networks (GANs) and Wasserstein Generative Adversarial Networks (WGANs) are utilized to generate synthetic data that closely mirrors actual patient data, thereby enhancing the robustness and generalizability of predictive tools for clinical application. These approaches also help address concerns related to data privacy and ethical considerations.

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), introduced by Goodfellow *et al.* [36], are a powerful class of generative models widely used for creating synthetic data that closely mimics real-world data distributions. GANs consist of two neural networks: a generator *G* and a discriminator *D,* which are trained simultaneously through an adversarial process. As depicted in Figure 2, the generator's role is to produce synthetic data samples by transforming random noise, typically drawn from a prior distribution such as a Gaussian, into data that resembles the real dataset. The discriminator, on the other hand, acts as a binary classifier that attempts to distinguish between genuine data samples and those generated by the generator.

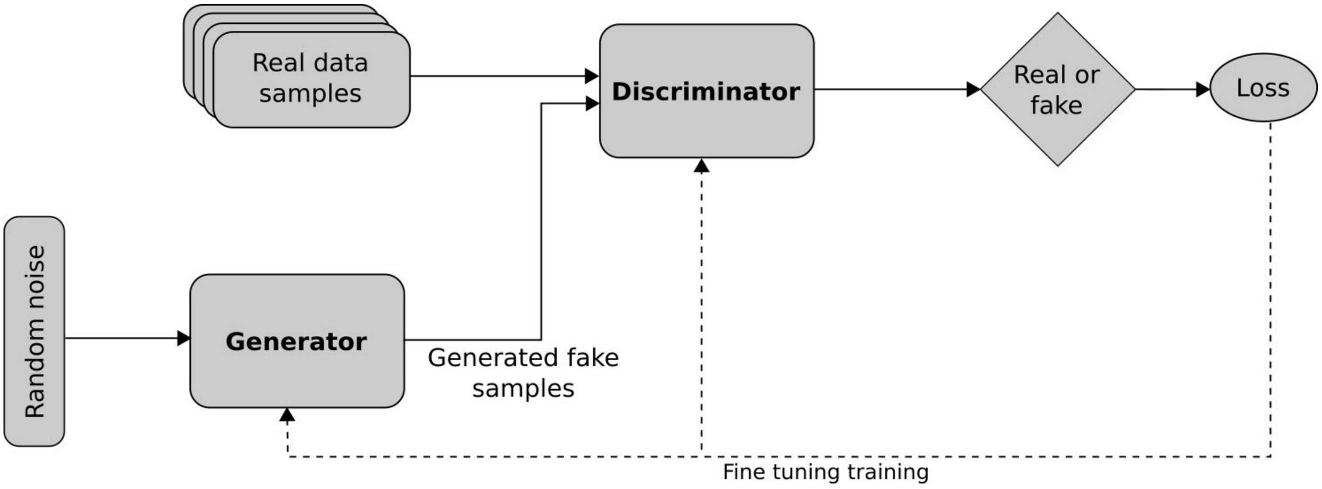


FIGURE 2. **Overview of the Generative Adversarial Network (GAN) Architecture.**

The adversarial training process of GANs is formulated as a minimax game, where the generator aims to "fool" the discriminator by producing data that the discriminator classifies as real, while the discriminator strives to correctly identify whether a given sample is real or fake. The objective function that governs this process is defined by Equation (3).

$$min_G max_D V(D,G) = E_{x \sim p_{data}(x)}[logD(x)] + E_{z \sim p_z(z)}[log(1 - D(G(z)))] \tag{3}$$

where:

- x represents real data samples drawn from the true data distribution $p_{data}(x)$.
- z is a latent variable (noise) sampled from a prior distribution $p_z(z)$.
- $G(z)$ denotes the synthetic sample generated by the generator from noise z.
- $D(x)$ provides the probability that a given input x is a real sample.

In this framework, the generator *G* learns to minimize the probability of the discriminator correctly identifying its synthetic outputs as fake, while the discriminator *D* aims to maximize the accuracy of its predictions. This continuous feedback loop drives both networks to improve iteratively: the generator becomes better at creating realistic data, and the discriminator becomes more adept at detecting synthetic samples. The training reaches equilibrium when the discriminator cannot distinguish between real and generated data, outputting a probability close to 0.5 for both[36].

GANs rely on two critical components to function effectively:

1. Adversarial Loss: The loss function quantifies the performance of both the generator and the discriminator. For the discriminator, the loss function can be expressed as shown in Equation (4).

$$L_D = -(E_{x\sim p_{data}(x)}[log\ D(x)] + E_{z\sim p_z(z)}[log(1 - D(G(z)))]) \tag{4}$$

For the generator, the goal is to minimize the loss function as defined in Equation (5).

$$L_G = -E_{z\sim p_z(z)}[log\ D(G(z))] \tag{5}$$

2. Optimization and Stability: Training GANs can be challenging due to issues like mode collapse, where the generator produces limited diversity in outputs, and training instability caused by non-convergence or oscillations. Techniques such as using alternative loss functions (e.g., Wasserstein loss for WGANs [35]), careful tuning of hyperparameters, label smoothing, and architectural adjustments like batch normalization and dropout layers are employed to mitigate these problems [37].

Model design and implementation

To effectively generate synthetic datasets that capture the complex characteristics of pediatric diabetes data for this study, a GAN was carefully designed and implemented with specific architectural choices and training parameters. These were selected to achieve an optimal balance between model complexity, training stability, and the generation of high-quality synthetic data. Table 2 provides a detailed overview of the key architectural components and training parameters used in this study.

TABLE 2. **GAN Architecture and Training Parameters for Synthetic Data Generation.**

Component	Description
Generator	4 dense layers: Input (150 neurons), Hidden layers (512, 1024, 2048 neurons) with Leaky ReLU activations, Batch Normalization after each hidden layer, Output layer with tanh activation.
Discriminator	5 dense layers: Input layer, Hidden layers (2048, 1024, 512, 256 neurons) with Leaky ReLU activations, Dropout layers (rate = 0.4) after each hidden layer, Output layer with sigmoid activation.
Optimizer	RMSprop with learning rate of 0.0001 for both generator and discriminator.
Training epochs	100 epochs with a batch size of 64.
Noise addition	Gaussian noise added to inputs of discriminator and final synthetic dataset to maintain diversity.
Label smoothing	Applied to real samples (0.9) and fake samples (0.1) to improve training stability.
Post-processing	Descaling and adjustment of discrete variables to ensure realistic data values.

The architecture and training parameters of the GAN model were selected based on established practices in the literature and tailored to meet the requirements of this study. The generator and discriminator were designed with dense layers and activation functions, such as Tanh and Leaky ReLU, which are widely recognized for their ability to improve gradient flow and stabilize adversarial training[36][38]. The RMSprop optimizer, combined with a carefully chosen learning rate, was employed to ensure stable convergence during training[37]. To enhance model robustness and diversity in the generated data, additional techniques such as batch normalization, noise addition, label smoothing, and dropout were integrated. These approaches, supported by prior research, have proven effective in improving training stability and preventing overfitting in GAN architectures[39][40]. Post-processing steps, including descaling and adjustment of discrete variables, were applied to ensure the synthetic datasets reflected realistic and clinically meaningful values. Collectively, these design choices contributed to the successful generation of synthetic datasets that accurately capture the complex characteristics of pediatric diabetes data.

Wasserstein Generative Adversarial Networks (WGANs)

Wasserstein Generative Adversarial Networks (WGANs) are an improved variant of traditional GANs, designed to address common challenges such as instability during training and mode collapse, where the generator fails to capture the full diversity of the real data distribution. Introduced by Arjovsky *et al.*[35], WGANs use the Wasserstein distance (also known as Earth Mover's Distance) as a loss function, which provides more stable gradients for training and leads to improved convergence properties. The Wasserstein distance is defined as shown in Equation (6).

$$W(p_r, p_g) = \ inf_{\gamma \epsilon \Pi(p_r, p_g)} E_{(x,y)\sim\gamma}[||x - y||] \tag{6}$$

where:

- p_r is the distribution of the real data.
- p_g is the distribution of the generated data by the generator G.
- $\Pi_{(p_r p_g)}$ represents the set of all joint distributions $\gamma(x,y)$ whose marginals are p_r and p_g, respectively.

Using this distance allows WGANs to maintain meaningful gradients throughout the training process, even when there is limited overlap between the real and generated data distributions. Instead of the discriminator outputting a probability, as in traditional GANs, the critic in a WGAN outputs a real-valued score, which estimates the Wasserstein distance between the real and generated data distributions. The generator is trained to minimize this distance, while the critic is trained to maximize it. This formulation leads to more stable training, as it avoids problems such as vanishing gradients that can occur with other types of loss functions[41].

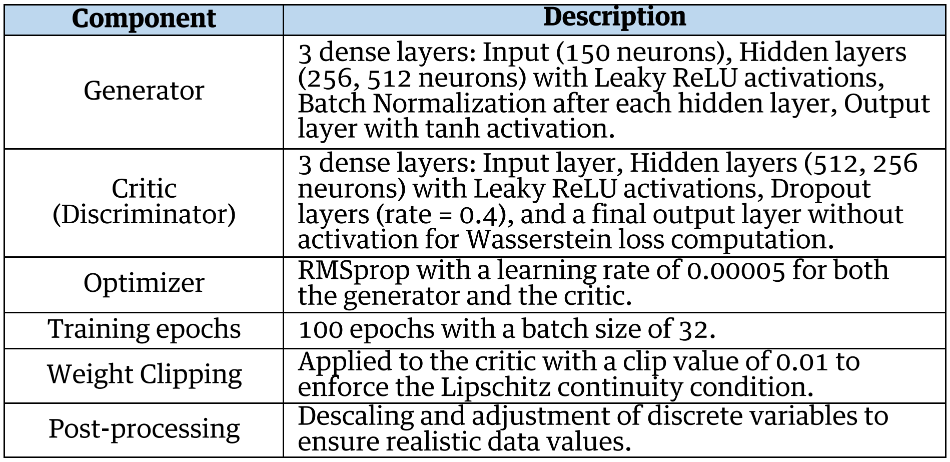
To ensure that the Wasserstein distance is valid, the critic function must satisfy the Lipschitz continuity condition, which is enforced by clipping the weights of the critic to a small range (e.g., [−0.01,0.01]). This weight clipping helps maintain stable gradients and promotes effective training, further reducing the likelihood of mode collapse and enhancing the quality of the synthetic data produced.

While WGANs introduce significant improvements in training dynamics through these changes, their overall architecture in terms of the core components (generator and critic) remains similar to that of a traditional GAN. Both models consist of a generator network that creates synthetic data and a discriminator (or critic in the case of WGAN) that evaluates the data. Therefore, the differences between GAN and WGAN primarily lie in the loss function and training strategies, rather than in the architectural layout itself [41][42]. For this reason, a separate figure illustrating the architecture of WGAN is not included, as it would visually mirror the figure already provided for the standard GAN.

Model design and implementation

To effectively generate synthetic datasets that capture the complex characteristics of pediatric diabetes data for this study, a WGAN was carefully designed and implemented with specific architectural choices and training parameters. The chosen architecture and parameters aim to leverage the advantages of the Wasserstein distance, ensuring stable training and high-quality data generation. The details are shown in Table 3.

TABLE 3. **WGAN Architecture and Training Parameters for Synthetic Data Generation.**



Component	Description
Generator	3 dense layers: Input (150 neurons), Hidden layers (256, 512 neurons) with Leaky ReLU activations, Batch Normalization after each hidden layer, Output layer with tanh activation.
Critic (Discriminator)	3 dense layers: Input layer, Hidden layers (512, 256 neurons) with Leaky ReLU activations, Dropout layers (rate = 0.4), and a final output layer without activation for Wasserstein loss computation.
Optimizer	RMSprop with a learning rate of 0.00005 for both the generator and the critic.
Training epochs	100 epochs with a batch size of 32.
Weight Clipping	Applied to the critic with a clip value of 0.01 to enforce the Lipschitz continuity condition.
Post-processing	Descaling and adjustment of discrete variables to ensure realistic data values.

The choice of these parameters and configurations reflects a balance between model complexity, training stability, and the ability to generate high-quality synthetic data. The generator and critic architectures, with three dense layers and Leaky ReLU activations, are specifically designed to capture the intricate patterns in pediatric diabetes data. Batch normalization is incorporated in the generator to stabilize gradient updates, while dropout is applied in the critic to prevent overfitting during adversarial training[37][41].

The RMSprop optimizer with a learning rate of 0.00005 was selected to ensure gradual convergence and prevent instability during the adversarial training process. This is particularly important in the WGAN framework, as the Wasserstein distance requires precise optimization to maintain the Lipschitz continuity condition[35]. Weight clipping with a small range (0.01) enforces this condition, further enhancing the stability and effectiveness of the training process[14].

Additionally, the number of training epochs (100) and the batch size (32) were carefully chosen to balance sufficient training iterations and computational efficiency. These configurations allow both the generator and critic to achieve optimal performance without overfitting the data. Post-processing adjustments, including descaling and the refinement of discrete variable values, ensure that the generated synthetic data remains realistic and clinically meaningful[38][39].

The WGAN framework was chosen for its ability to address limitations in traditional GANs, particularly in stabilizing the training process and generating diverse, high-quality data. By combining the standard GAN approach with the more robust WGAN framework, this study leverages the strengths of each model to produce synthetic datasets that are not only accurate but also diverse and suitable for developing reliable predictive models.

Evaluation metrics

To rigorously assess the quality of the synthetic datasets generated by the GAN and WGAN models, three complementary evaluation methods are employed: Jensen-Shannon Divergence (JSD), Kullback-Leibler Divergence (KLD), and Predictive Model Evaluation through a Random Forest classifier. These metrics provide a comprehensive analysis of the fidelity, statistical alignment, and utility of the synthetic data by quantifying distributional similarities and assessing their practical application in predictive tasks.

Jensen-Shannon Divergence (JSD)

The Jensen-Shannon Divergence (JSD) is a statistical measure used to quantify the similarity between two probability distributions[43]. As a symmetrized and smoothed version of the Kullback-Leibler divergence, the JSD provides a bounded and interpretable metric for assessing how one distribution diverges from another. It is particularly suitable for evaluating the performance of generative models, such as GANs and WGANs, by measuring the extent to which the synthetic data replicates the distributional characteristics of real-world data[44][45].

Mathematically, the JSD between two discrete probability distributions *P* and *Q* is defined as shown in Equation (7):

$$JSD(P \mid\mid Q) = \frac{1}{2}KL(P \mid\mid M) + \frac{1}{2}KL(Q \mid\mid M) \tag{7}$$

where:

- *P* and *Q* are the two distributions being compared.
- *M=½(P+Q)* is the average distribution.
- *KL(P || M)* and *KL(Q || M)* ar re the Kullback-Leibler divergences between each distribution and the mean distribution *M*.

The JSD is symmetric *(JSD(P || Q) = JSD(Q || P))* and is bounded between 0 and 1. A value of 0 indicates that the two distributions are identical, while values closer to 1 indicate greater divergence. This makes JSD a robust metric for comparing distributions, as it remains finite and interpretable even when the distributions have non-overlapping support.

In this study, the JSD is employed to rigorously evaluate the similarity between the real pediatric diabetes dataset and the synthetic datasets generated by the GAN and WGAN models. The JSD is chosen because it provides a comprehensive measure of distributional similarity, capturing both general alignment and specific differences between the real and synthetic data distributions. This is particularly important given the complex nature of pediatric diabetes data, which may contain nuanced patterns and variability that must be accurately reflected in the synthetic data.

The implementation of the Jensen-Shannon Divergence (JSD) in this study is carried out through the following steps:

1. Data Alignment: Both the real and synthetic datasets are aligned to ensure they contain identical feature columns. Any missing values are removed to maintain consistency and comparability across all features.
2. Feature-wise Calculation of JSD: For each feature in the datasets, probability distributions are represented by constructing histograms using a consistent binning strategy. The JSD is computed for each feature by comparing the normalized histograms of real and synthetic data, providing a divergence measure that reflects the degree of similarity between the synthetic and real data distributions for that specific feature.
3. Aggregation Across Features: The JSD values calculated for all features are averaged to produce a single, overall measure of divergence. This mean JSD value captures the overall similarity between the synthetic and real datasets, considering all features simultaneously.

The use of JSD in this study is justified by its capacity to provide a clear and quantitative assessment of the extent to which the synthetic data replicates the distributional characteristics of the original data. By evaluating both global and feature-specific similarities, the JSD offers a robust measure of the performance of the generative models, ensuring that the synthetic datasets produced are suitable for further analysis and modeling within the context of pediatric diabetes research.

Kullback-Leibler Divergence (KLD)

The Kullback-Leibler Divergence (KLD) is a fundamental measure from information theory that quantifies the difference between two probability distributions[17]. It is commonly used to evaluate the performance of generative models by measuring the "cost" or "loss of information" when approximating one probability distribution with another[36][46]. In this study, KLD is employed to assess how closely the synthetic datasets generated by the GAN and WGAN models match the distribution of the real pediatric diabetes data.

Mathematically, KLD between two discrete probability distributions P (real data) and Q (synthetic data) is defined as shown in Equation (8):

$$KL(P\ ||\ Q) = \sum_{i} P(i) log(\frac{P(i)}{Q(i)}) \tag{8}$$

where:

- *P(i)* represents the probability of event *i* in the real data distribution.
- *Q(i)* represents the probability of event *i* in the synthetic data distribution.

In this study, both JSD and KLD are used to provide a comprehensive evaluation of the synthetic data quality. The combination of these two metrics offers a robust framework for assessing how effectively the generative models replicate the distributional characteristics of the real pediatric diabetes data. While JSD provides a global measure of similarity that is symmetric and bounded, making it suitable for general assessments of distributional overlap, KLD offers a more detailed perspective by identifying specific divergences where the synthetic data may differ significantly from the real data[47][48].

The implementation of KLD follows these steps:

1. Data Preparation: Both the real and synthetic datasets are aligned to ensure they have matching feature

sets, with missing values removed to enable accurate comparison.

2. Feature-wise KLD Calculation: For each feature, probability distributions are estimated by constructing normalized histograms. KLD is computed for each feature by comparing the real and synthetic distributions, providing insight into specific areas where the synthetic data may diverge from the real data.
3. Aggregation Across Features: The KLD values for all features are aggregated to derive an overall measure of divergence. This overall KLD score reflects the general fidelity of the synthetic dataset in representing the real data, considering all features collectively.

The use of both JSD and KLD in this study is justified by their complementary strengths. JSD is robust to extreme discrepancies and provides an overall measure of similarity, making it useful for confirming that the synthetic data generally aligns with the real data. KLD, on the other hand, offers a more granular assessment, pinpointing exact areas of divergence, which is crucial for refining the generative models. Together, these metrics ensure that the synthetic datasets generated are not only globally similar to the real data but also accurately capture its detailed characteristics, thereby validating their use for subsequent analysis and predictive modeling in pediatric diabetes research.

Predictive Model Evaluation

The quality of synthetic datasets is not only defined by their statistical alignment with real data but also by their practical utility in downstream machine learning tasks. To this end, a predictive model evaluation framework is incorporated as an additional metric for assessing the efficacy of the generated synthetic data. This method evaluates how well synthetic datasets support classification tasks, serving as a proxy for their utility in real-world applications.

The Random Forest algorithm is selected as the predictive model due to its inherent robustness, capacity to handle high-dimensional data, and ability to model complex feature interactions[49][50]. Random Forest operates as an ensemble of decision trees, where each tree is constructed using a bootstrap sample of the training data. At each decision node, a random subset of features is considered for splitting, introducing variability that enhances the generalizability of the model. The final prediction is obtained through majority voting (classification) or averaging (regression) across all trees, ensuring resilience to overfitting[51].

Mathematically, the Random Forest Classifier constructs T decision trees, where each tree $t \in T$ minimizes the Gini impurity I_g for classification tasks, as shown in Equation (9).

$$I_g = \sum_{i=1}^{C} p_i(1 - p_i) \tag{9}$$

Here, C denotes the number of classes, and p_i, is the proportion of samples belonging to class i in the current node. This criterion guides the selection of optimal splits at each node, ensuring that the trees effectively capture the underlying patterns in the data[52].

By employing Random Forest, this study introduces a pragmatic approach to assess the synthetic data's relevance in predictive modeling. This framework complements statistical measures like Jensen-Shannon Divergence (JSD) and Kullback-Leibler Divergence (KLD) by providing a task-oriented perspective, ensuring a comprehensive evaluation of synthetic data quality.

RESULTS AND DISCUSSION

This section presents the findings from the experimental processes applied in this study and discusses their implications in the context of pediatric diabetes research. The analysis begins with the outcomes of data preprocessing steps, including feature filtering, management of data incompleteness, and data normalization. It then describes the application of generative models (GANs and WGANs) to create synthetic datasets. The quality and fidelity of these generated datasets are subsequently assessed using the Jensen-Shannon Divergence (JSD), Kullback-Leibler Divergence (KLD), and Predictive Model Evaluation metrics. Together, these approaches provide a comprehensive evaluation of the synthetic data, offering insights into their distributional similarity to real-world data and their utility in downstream machine learning tasks. This combination ensures a robust assessment of their potential applicability for future research.

Data Preprocessing

Feature filtering

The preprocessing of the dataset involves several critical steps to ensure its suitability for synthetic data generation and subsequent analysis. The original dataset comprises 834 records with 22 features, including various demographic, clinical, and biochemical variables relevant to pediatric diabetes research. To improve the dataset's quality and focus, three features are removed due to their lack of direct relevance to the study objectives, resulting in a refined dataset containing 19 features. The excluded features are detailed in Table 4.

TABLE 4. **Features Removed During Preprocessing.**

Feature Name	Reason for removal
Location	Lack of direct relevance to study goals
StudyProject	
RecordID	

Data Incompleteness

Handling data incompleteness presents a substantial challenge in the preprocessing stage. Among the 834 records, 462 (approximately 55.4 %) exhibit at least one missing value. To address this issue, an imputation strategy is employed, wherein missing values are replaced with the mean of their respective features. This approach maintains the dataset's integrity by preserving the total number of records and minimizing potential biases that could arise from data omission. Following this imputation process, the dataset retains all 834 records, as summarized in Table 5.

TABLE 5. **Summary of Handling Missing Data.**

Metric	Value
Total number of records in the original dataset	834
Records with at least one missing value	465 (55.4 %)
Total number of records after imputation	834

These preprocessing steps are vital in preparing the dataset for further analysis. They ensure that the data is clean, complete, and properly scaled, thereby enhancing the reliability and quality of the synthetic data generated.

Data normalization

After the imputation of missing values, a normalization process is applied to ensure consistency in the scale of the dataset's features, which is essential for the effective training of generative models. A Min-Max normalization technique is utilized, scaling each feature to a range between 0 and 1. This standardization mitigates the risk of features with larger numerical ranges disproportionately influencing the model training process, thereby ensuring that all features contribute equitably to the generative modeling efforts.

Synthetic data generation and assessment

While traditional data augmentation (DA) techniques have proven effective in domains such as images or text, where transformations like rotations or interpolations preserve the semantics of the original data, their application to clinical numerical data presents significant limitations. These techniques often fail to adequately capture the complex relationships between variables or maintain the clinical coherence required in healthcare datasets. For instance, Shorten and Khoshgoftaar[53] highlight how DA techniques can enhance model generalization in image classification but note their limited applicability to more structured and sensitive domains. Similarly, Choi *et al.*[54] demonstrate that GANs outperform DA in generating tabular clinical data by better preserving multidimensional relationships and ensuring consistency.

To address these challenges, this study employs Generative Adversarial Networks (GANs) and Wasserstein GANs (WGANs), which are better suited to model multidimensional distributions and generate synthetic data with higher fidelity and clinical consistency. This approach not only replicates the statistical properties of the original data but also preserves intrinsic relationships between features, a critical requirement in clinical research.

The use of Generative Adversarial Networks (GANs) and Wasserstein Generative Adversarial Networks (WGANs) is crucial for generating synthetic datasets that closely replicate the distributional characteristics of real pediatric diabetes data. Each model's effectiveness in producing realistic synthetic data is assessed based on its ability to preserve the statistical properties of the original dataset and its utility in downstream machine learning tasks. To achieve this, the quality of the synthetic data is evaluated using three complementary metrics: Jensen-Shannon Divergence (JSD), Kullback-Leibler Divergence (KLD), and the performance of a supervised classification model. Together, these methods provide a rigorous assessment of the statistical alignment and practical applicability of the synthetic datasets, ensuring their relevance for research and predictive modeling.

Evaluation of GAN-Generated Synthetic Data

The Generative Adversarial Network (GAN) is utilized to create a synthetic dataset designed to closely replicate the distributional characteristics of the original pediatric diabetes data. A synthetic dataset consisting of 417 records, representing 50 % of the original dataset's size, is generated to maintain a balance between capturing the complexity of the data and ensuring a robust comparison between the real and synthetic datasets.

The effectiveness of the GAN in generating realistic synthetic data is evaluated through a comparative analysis of the global distributions of the real and GAN-generated data. Figure 3 illustrates this comparison by aggregating the normalized values of all features into a single composite measure for each record, providing a comprehensive view of the similarity between the two datasets. This composite measure summarizes the overall alignment between the datasets, capturing both the central tendencies and variability of the original data.

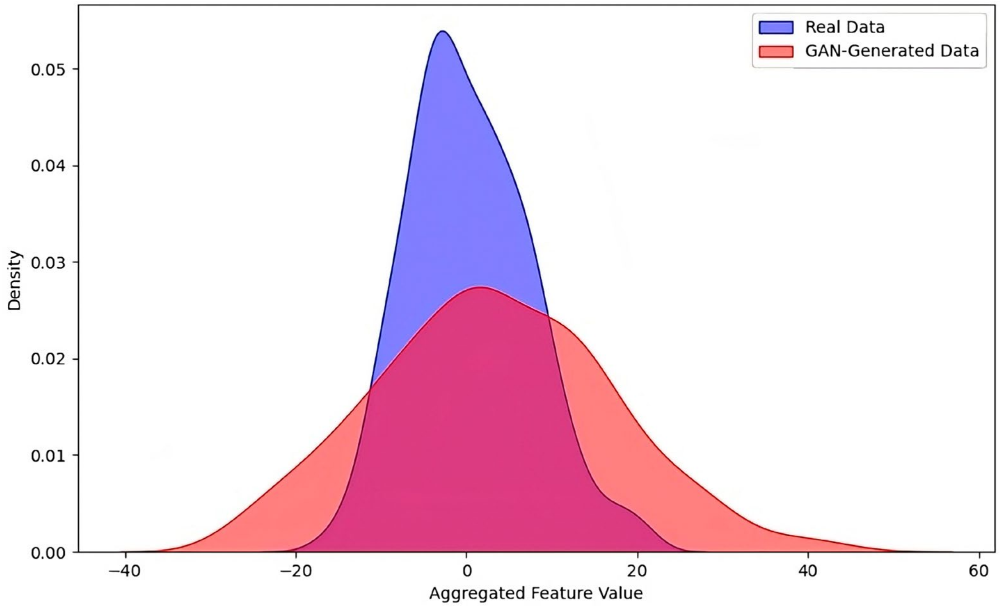


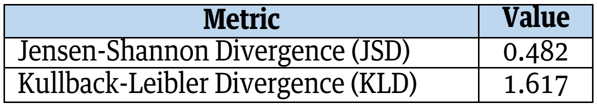
FIGURE 3. **Global Distribution Comparison Between Real and GAN-Generated Data.**

In Figure 3, the "Aggregated Feature Value" represents a combined metric derived by summing all normalized feature values for each record, while the probability density indicates the relative frequency of these aggregated values across the dataset. The curve corresponding to the real dataset is depicted in blue, while the red curve represents the GAN-generated dataset. The visualization demonstrates that the synthetic data generated by the GAN aligns well with the general shape of the real data distribution, capturing key patterns and trends, particularly in the central regions where the curves overlap closely.

Discrepancies are observed in the tails of the distribution, where the GAN fails to replicate less frequent and extreme values. This limitation arises from the training instability commonly associated with traditional GAN architectures, particularly when modeling complex, high-dimensional data. While the GAN successfully reproduces the dominant statistical features of the real dataset, these deviations highlight its reduced ability to capture rare patterns that require more stable and fine-tuned training processes.

To further quantify the similarity between the real and GAN-generated data, two divergence metrics are employed: the Jensen-Shannon Divergence (JSD) and the Kullback-Leibler Divergence (KLD). The values of these metrics, presented in Table 6, offer a detailed view of the effectiveness of the GAN in approximating the real data.

TABLE 6. **Divergence Metrics for GAN-Generated Synthetic Data.**



Metric	Value
Jensen-Shannon Divergence (JSD)	0.482
Kullback-Leibler Divergence (KLD)	1.617

As shown in Table 6, the Jensen-Shannon Divergence (JSD) value of 0.482 indicates a substantial degree of similarity between the distributions of the real and GAN-generated data. The relatively low JSD suggests that the GAN effectively captures the overall structure and dominant statistical features of the original dataset. This result confirms that the global patterns, including central tendencies and frequent values, are well-represented in the synthetic data.

TABLE 7. **Classification Performance Comparison between Real and GAN-Generated Synthetic Data.**

Dataset	Accuracy	Precision	Recall	F1 Score	ROC AUC
Real data	0.98	0.97	0.97	0.97	0.98
Synthetic data	0.94	0.93	0.93	0.93	0.94

The Kullback-Leibler Divergence (KLD) value of 1.617 provides additional insight into the alignment of the two distributions. While this value remains relatively low, it is slightly higher than the JSD, indicating that discrepancies are more pronounced in regions with lower probabilities, such as the tails of the distribution. This reflects the GAN's difficulty in modeling less frequent or extreme values, which aligns with the limitations observed visually in Figure 3. Together, these divergence metrics demonstrate that while the GAN-generated data closely approximates the real data, there is room for improvement in capturing rare patterns and achieving a more precise representation of the complete data distribution.

To complement this evaluation, the quality of the GAN-generated synthetic data is also assessed through its utility in a supervised learning context. A Random Forest classifier is trained separately on the real dataset and the GAN-generated synthetic dataset, using a binary classification task where the target variable indicates the presence or absence of diabetes. The feature Insulin (labeled as Insulin_ugml in the dataset) is used to define this target variable, as insulin levels are clinically recognized as a key indicator for diabetes diagnosis. Based on medical thresholds, records with insulin levels greater than 25 μU/mL are labeled as diabetic (1), while others are labeled as non-diabetic (0).

Key metrics, including accuracy, precision, recall, F1-score, and ROC AUC, are computed to evaluate the performance of the model in each case. The results, summarized in Table 7, provide a point of comparison to understand how well the synthetic data supports predictive modeling relative to the real data.

The performance metrics in Table 7 show that the GAN-generated synthetic data achieves strong and consistent results across all evaluated metrics. Although a slight drop of approximately 4 % is observed in accuracy, precision, recall, and F1-score compared to the real data, this difference is within acceptable limits for synthetic datasets. The marginal decline reflects the challenge of fully replicating the nuanced variability of real-world data, particularly in capturing less frequent patterns or edge cases.

The relatively high precision and recall values indicate that the synthetic data retains the critical predictive features necessary for identifying positive cases while effectively minimizing false positives. This balance is essential for clinical applications, where both sensitivity (recall) and specificity (precision) are critical for reliable predictions. Furthermore, the high ROC AUC score confirms that the model trained on GAN-generated data maintains a strong discriminative capability, demonstrating its ability to distinguish between classes effectively.

These findings highlight that, despite minor discrepancies, the GAN-generated data successfully preserves the underlying relationships and predictive features of the real dataset. This validates the GAN's practical utility for training machine learning models in scenarios where real data are scarce or access is restricted, offering a robust alternative for enhancing predictive modeling in pediatric diabetes research.

However, while these results are promising, they also reflect the inherent challenges of working with pediatric data, which tend to be highly variable due to the ongoing growth and development of children. The diversity in growth rates, physiological changes, and individual responses to health conditions can result in complex patterns that are difficult to replicate fully with synthetic models. These characteristics contribute to the observed differences between the real and synthetic data and underscore the need for careful consideration when generating synthetic datasets in this context.

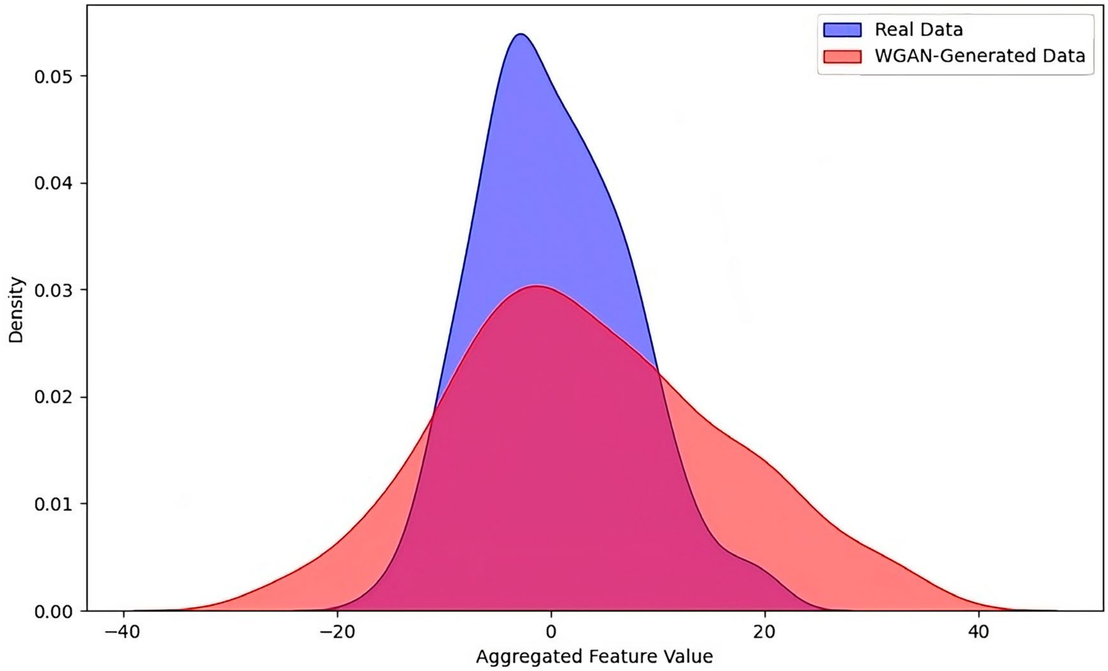


FIGURE 4. **Global Distribution Comparison Between Real and WGAN-Generated Data.**

Overall, the combined analysis of the visual comparison, divergence metrics, and supervised learning performance suggests that the GAN is highly effective in generating synthetic data that not only closely matches the general distribution of the real data but also retains the critical predictive features necessary for accurate classification. The comparable performance of the Random Forest classifier on both the real and synthetic datasets reinforces the validity of the synthetic data, demonstrating that it encapsulates the essential patterns and relationships required for supervised learning tasks. This comprehensive evaluation highlights the potential of GAN-generated data to serve as a reliable alternative for training machine learning models when real data is scarce or inaccessible.

Evaluation of WGAN-Generated Synthetic Data

The Wasserstein Generative Adversarial Network (WGAN) is the second method used for generating synthetic data in this study, aimed at producing a dataset that closely replicates the characteristics of the original pediatric diabetes data. The WGAN model generates a synthetic dataset of 417 records, mirroring the size used in the GAN-generated dataset to maintain consistency in comparisons.

The performance of the WGAN is evaluated by comparing the global distributions of the real and WGAN-generated data, as shown in Figure 4. As in the GAN analysis, the normalized values of all features are aggregated into a single composite measure for each record, allowing a direct comparison of the overall alignment between the datasets.

TABLE 8. **Divergence Metrics for WGAN-Generated Synthetic Data.**

Metric	Value
Jensen-Shannon Divergence (JSD)	0.480
Kullback-Leibler Divergence (KLD)	1.584

This method provides a comprehensive view of how well the WGAN captures the underlying structure and variability of the real data.

In Figure 4, the "Aggregated Feature Value" represents a combined metric derived by summing all normalized feature values for each record, while the probability density indicates the relative frequency of these aggregated values across the dataset. The curve corresponding to the real dataset is depicted in blue, while the red curve represents the WGAN-generated dataset. The visualization shows that the WGAN-generated data aligns closely with the real data distribution, particularly around the central peak and mid-range values, where the curves exhibit almost perfect overlap. This alignment indicates that the WGAN successfully captures the most frequent and dominant statistical patterns in the data, closely approximating the global structure observed in the real dataset.

Compared to the traditional GAN, the WGAN demonstrates a significant reduction in discrepancies at the tails of the distribution, where less frequent and extreme values occur. This improvement reflects the enhanced stability of the WGAN training process, which allows it to model complex patterns with greater accuracy. By reducing deviations in the tails, the WGAN effectively retains both dominant and rare features, preserving the variability essential for realistic synthetic data generation. This capacity makes the WGAN-generated data not only a reliable representation of the real dataset but also a valuable resource for downstream machine learning tasks, particularly in scenarios with limited real data availability.

TABLE 9. **Classification Performance comparison between Real and WGAN-Generated Synthetic Data.**

Dataset	Accuracy	Precision	Recall	F1 Score	ROC AUC
Real data	0.98	0.97	0.97	0.97	0.98
Synthetic data	0.96	0.95	0.95	0.95	0.97

To further quantify the similarity between the real and WGAN-generated data, the same divergence metrics used for the GAN evaluation, the Jensen-Shannon Divergence (JSD) and the Kullback-Leibler Divergence (KLD), are calculated. The results, presented in Table 8, provide a detailed assessment of the WGAN's performance.

As shown in Table 8, the Jensen-Shannon Divergence (JSD) value of 0.480 is slightly lower than the value obtained with the GAN (0.482), indicating a marginal improvement in the similarity between the WGAN-generated and real data distributions. This lower JSD suggests that the WGAN more effectively captures the global statistical structure of the dataset, aligning closely with the dominant patterns observed in the real data.

The Kullback-Leibler Divergence (KLD) value of 1.584 is also marginally reduced compared to the GAN (1.617), highlighting a better representation of less frequent and extreme values in the tails of the distribution. This improvement reflects the enhanced stability of the WGAN training process, which enables it to model complex, low-probability patterns with greater accuracy. Collectively, these results demonstrate that the WGAN outperforms the traditional GAN in both global alignment and in capturing subtle variations, making it a more reliable approach for generating high-fidelity synthetic data.

Similar to the evaluation conducted for GAN, the quality of the WGAN-generated synthetic data is assessed through its utility in a supervised learning context. This analysis follows the same approach as the previous section, where a Random Forest classifier is trained on both the real dataset and the WGAN-generated synthetic dataset to predict the presence or absence of diabetes based on insulin levels. Key metrics such as accuracy, precision, recall, F1-score, and ROC AUC are calculated to evaluate the model's performance on both datasets. The results are summarized in Table 9.

As shown in Table 9, the Random Forest classifier trained on WGAN-generated synthetic data achieves performance metrics that are closely aligned with those obtained using real data. The accuracy, precision, recall, F1-score, and ROC AUC for the WGAN-generated data show only a slight decrease of approximately 2 % compared to the real dataset. This minor drop highlights the ability of the WGAN to generate high-fidelity synthetic data that preserves critical predictive features and relationships.

When compared to the GAN-generated data (Table 7), the WGAN demonstrates improved performance across all metrics, particularly in recall and ROC AUC. The slight enhancements in recall indicate that the WGAN-generated data better captures edge cases and less frequent patterns, reducing false negatives more effectively. Additionally, the ROC AUC value of 0.97 for the WGAN surpasses the 0.94 obtained with the GAN, showcasing the superior discriminative power of the WGAN-generated data.

The WGAN’s ability to produce synthetic data with enhanced stability and improved representation of subtle and less frequent patterns demonstrates its superiority over traditional GANs. By addressing the limitations observed in earlier models, the WGAN-generated data offer a reliable and realistic alternative for training machine learning models. This is particularly valuable in pediatric research, where access to real-world datasets is often limited, and the quality of synthetic data plays a crucial role in enabling robust predictive modeling.

The results obtained in this study, particularly the superior performance of Wasserstein Generative Adversarial Networks (WGANs) compared to traditional GANs, align with findings from other research on synthetic data generation in healthcare. Goncalves *et al.*[55] conducted a thorough examination of synthetic data generation techniques for cancer patient data, focusing on methods such as probabilistic models, Bayesian networks, and GANs. Their results demonstrated the potential of these techniques to approximate real datasets, achieving a Kullback-Leibler (KL) divergence score of 0.47 for cancer datasets, which is comparable to the results obtained in this study for pediatric diabetes data (KL divergence of 1.584 for WGAN-generated data). Although the divergence in this study is slightly higher, it is crucial to consider the unique challenges posed by pediatric data. Growth and developmental variability introduce significant complexity in modeling, making pediatric data inherently more difficult to capture accurately.

Other studies in synthetic healthcare data generation also reflect similar trends, particularly in the challenges of handling complex datasets. Xu *et al.*[56] applied Conditional GANs (CTGAN) to generate synthetic tabular data and found that CTGAN outperforms Bayesian networks and other GAN-based models on several real-world datasets. Although their focus was not on pediatric data, their results highlight the effectiveness of using conditional GANs for generating structured data patterns. However, their work, like many others, focuses primarily on adult datasets, where the variability present in pediatric populations is not a significant factor. This difference underscores the need for models capable of adapting to more dynamic physiological conditions, such as those found in children with diabetes, where growth and development create a more challenging environment for accurate synthetic data generation.

While the generation of synthetic healthcare data has advanced significantly, much of the focus has been on adult datasets, with limited attention paid to pediatric populations. This study addresses a significant gap in the literature by applying advanced generative models, specifically GANs and WGANs, to pediatric diabetes data. Pediatric datasets present unique challenges due to the high variability in physiological characteristics as children grow and develop. This inherent variability complicates the modeling process and requires more sophisticated approaches to accurately reflect the underlying data distributions. To address these complexities, WGANs were employed due to their stability in training and ability to handle the nuanced data patterns found in pediatric populations. Moreover, the use of Jensen-Shannon Divergence (JSD) and Kullback-Leibler Divergence (KLD) as evaluation metrics ensures that the generated datasets maintain the essential statistical properties necessary for predictive modeling.

To complement the evaluation of these divergence metrics, the quality of the WGAN-generated synthetic data was further assessed through its utility in a supervised learning context. A Random Forest classifier, trained on both real and WGAN-generated synthetic data, revealed that the synthetic data maintained essential predictive features, with classification performance metrics (accuracy, precision, recall, F1 score, and ROC AUC) comparable to the real data. These results demonstrate that the WGAN-generated data can be effectively used in predictive models for pediatric diabetes, showing that it not only captures the global distribution of the real data but also retains the critical features required for accurate classification.

By generating high-quality synthetic datasets for pediatric diabetes, this work can significantly enhance predictive models and personalized treatment approaches, addressing the critical scarcity of real-world pediatric datasets in this domain. To our knowledge, no studies to date have applied generative models specifically to pediatric data in the context of diabetes research, making this contribution particularly novel.

Limitations and adaptability

While this study demonstrates the successful use of GANs and WGANs to generate high-quality synthetic datasets for pediatric diabetes research, certain limitations must be acknowledged. The quality and representativeness of the generated synthetic data depend heavily on the characteristics of the input dataset. Small sample sizes, incomplete data, or datasets with imbalanced classes may hinder the models' ability to capture rare or subtle patterns accurately. This limitation is particularly significant in pediatric datasets, where inherent variability due to growth, development, and physiological changes introduces additional challenges. Ensuring high-quality input data remains a critical prerequisite for achieving reliable synthetic data generation.

Another limitation relates to the specificity of the current methodology to pediatric diabetes data. Although GANs and WGANs are generalizable techniques, their application to other datasets or domains requires adaptation. For instance, datasets with different structures, such as high-dimensional multi-omics data or longitudinal patient records, may require modifications to the preprocessing pipeline, including tailored strategies for handling missing data, normalization techniques, or feature selection. Likewise, adjustments to the model architectures may be necessary to optimize performance for datasets with greater complexity or variability.

Despite these limitations, the proposed methodology provides a solid foundation for generating synthetic datasets in healthcare and beyond. Future research can explore its adaptability to other medical conditions, larger datasets, and alternative domains, such as genomics or environmental monitoring, where data scarcity is also a significant challenge. Additionally, further evaluation across diverse datasets could validate the generalizability of this approach, ensuring its broader applicability in machine learning-driven predictive modeling.

CONCLUSIONS

This study explored the use of Generative Adversarial Networks (GANs) and Wasserstein GANs (WGANs) to generate synthetic datasets for pediatric diabetes research, addressing critical data scarcity in this domain. The dual approach combining statistical evaluation metrics and supervised classification validated the fidelity and practical utility of the synthetic data. By employing Jensen-Shannon Divergence (JSD) and Kullback-Leibler Divergence (KLD), the statistical alignment of synthetic datasets with real data was confirmed. Specifically, the GAN achieved a JSD of 0.482 and a KLD of 1.617, while the WGAN demonstrated superior performance with a JSD of 0.480 and a KLD of 1.584, indicating improved stability and fidelity in capturing complex data patterns.

The use of a Random Forest classifier to assess performance on classification tasks further demonstrated the capacity of synthetic datasets to retain predictive features essential for machine learning applications. The model trained on real data achieved an accuracy of 0.98, while the GAN-generated data attained 0.94, and the WGAN-generated data achieved 0.96, with similarly high scores for precision, recall, F1 score, and ROC AUC. These results highlight the WGAN's enhanced ability to approximate real dataset performance and reinforce its suitability for predictive modeling.

The results revealed that both GAN and WGAN methodologies successfully captured the complex distribution of pediatric diabetes data, with WGANs offering enhanced stability and improved fidelity. These findings are consistent with the theoretical advantages of the Wasserstein distance in mitigating common challenges in GAN training. Classification performance metrics further underscored the reliability of WGAN-generated data, which achieved metrics close to real data across accuracy, precision, recall, F1 score, and ROC AUC.

Beyond technical evaluations, this work contributes significantly by advancing the application of generative models in pediatric diabetes—a domain characterized by unique challenges, such as high variability in physiological characteristics and ethical constraints in data collection. Unlike prior studies that focused primarily on adult datasets or statistical metrics, this study introduces a comprehensive evaluation framework encompassing both distributional similarity and predictive utility.

Future work should aim to refine generative architectures and incorporate domain-specific knowledge to address the inherent complexities of pediatric datasets more effectively. Potential directions include exploring alternative loss functions, integrating temporal and longitudinal data, and developing hybrid models that merge generative techniques with rule-based systems. Expanding these methodologies to other pediatric conditions could further establish their relevance in healthcare research, paving the way for improved diagnostic tools and personalized treatment strategies.

This study demonstrates the transformative potential of synthetic data in overcoming real-world constraints, enhancing AI-driven solutions for pediatric diabetes management, and supporting broader applications in medical research. By rigorously validating synthetic datasets' fidelity and usability, this work lays the foundation for more robust and generalizable machine learning applications in healthcare.

ETHICAL STATEMENT

The study protocol was approved by the Ethics Committee of the Mexican Social Security Institute (approval number: R-2016-785-097), in compliance with the guidelines set forth by the National Bioethics Commission (CONBIOETICA-09-CEI-009-20160601).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

A. G.-D. conceptualization, methodology, supervision, writing - original draft, validation, and funding acquisition; C. E. G.-T. conceptualization, software, visualization, methodology, writing - review & editing, and validation; R. M.-Q. conceptualization, data curation, formal analysis, investigation, writing - review & editing, and validation; M. C.-L. investigation, resources, validation, and data curation; M. A. V.-M. formal analysis, visualization, writing - review & Editing, and validation; E. A.-C. project administration, resources, writing - review & Editing.

REFERENCIAS

[1] D. J. G. Carrasco Ramírez, M. Islam, and I. H. Even, "Machine Learning Applications in Healthcare: Current Trends and Future Prospects," J. Artif. Intell. Gen. Sci., vol. 1, no. 1, Art. no. 1, 2024, doi: https://doi.org/10.60087/jaigs.v1i1.33

[2] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," IEEE Rev. Biomed. Eng., vol. 14, pp. 156-180, 2021, doi: https://doi.org/10.1109/RBME.2020.3013489

[3] M. Javaid et al., "Significance of machine learning in healthcare: Features, pillars and applications," Int. J. Intell. Netw., vol. 3, pp. 58-73, 2022, doi: https://doi.org/10.1016/j.ijin.2022.05.002

[4] A. García-Domínguez et al., "Optimizing Clinical Diabetes Diagnosis through Generative Adversarial Networks: Evaluation and Validation," Diseases, vol. 11, no. 4, 2023, art. no. 134, doi: https://doi.org/10.3390/diseases11040134

[5] A. García-Domínguez et al., “Diabetes Detection Models in Mexican Patients by Combining Machine Learning Algorithms and Feature Selection Techniques for Clinical and Paraclinical Attributes: A Comparative Evaluation,” J. Diabetes Res., vol. 2023, 2023, art. no. 9713905, doi: https://doi.org/10.1155/2023/9713905

[6] J. G. Elmore and C. I. Lee, “Data Quality, Data Sharing, and Moving Artificial Intelligence Forward,” JAMA Netw. Open, vol. 4, no. 8, 2021, art. no. e2119345, doi: https://doi.org/10.1001/jamanetworkopen.2021.19345

[7] H. Sun et al., “IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045,” Diabetes Res. Clin. Pract., vol. 183, 2022, art. no. 109119, doi: https://doi.org/10.1016/j.diabres.2021.109119

[8] C. W. Tsao et al., “Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association,” Circulation, vol. 145, no. 8, pp. e153-e639, 2022, doi: https://doi.org/10.1161/cir.0000000000001052

[9] L. C. Stene and J. Tuomilehto, “Epidemiology of Type 1 Diabetes,” in Textbook of Diabetes, R.I.G. Holt and A. Flyvbjerg (Eds.)., Pondicherry, India: Wiley, 2024, ch. 4, pp. 41-54, doi: https://doi.org/10.1002/9781119697473.ch4

[10] G. Imperatore et al., “Projections of Type 1 and Type 2 Diabetes Burden in the U.S. Population Aged <20 Years Through 2050: Dynamic modeling of incidence, mortality, and population growth,” Diabetes Care, vol. 35, no. 12, pp. 2515-2520, 2012, doi: https://doi.org/10.2337/dc12-0669

[11] P. Bjornstad et al., “Youth-onset type 2 diabetes mellitus: an urgent challenge,” Nat. Rev. Nephrol., vol. 19, no. 3, pp. 168-184, 2023, doi: https://doi.org/10.1038/s41581-022-00645-1

[12] J. Jordon et al., “Synthetic Data -- what, why and how?,” 2022, arXiv:2205.03257, doi: https://doi.org/10.48550/arXiv.2205.03257

[13] V. C. Pezoulas et al., “Synthetic data generation methods in healthcare: A review on open-source tools and methods,” Comput. Struct. Biotechnol. J., vol. 23, pp. 2892-2910, 2024, doi: https://doi.org/10.1016/j.csbj.2024.07.005

[14] I. Goodfellow et al., “Generative Adversarial Networks,” 2014, arXiv:1406.2661, doi: https://doi.org/10.48550/arXiv.1406.2661

[15] M. Arjovsky, S. Chintala, L. Bottou, “Wasserstein Generative Adversarial Networks,” in Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017, pp. 214-223, doi: https://dl.acm.org/doi/10.5555/3305381.3305404

[16] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo, “The Jensen-Shannon divergence,” J. Frankl. Inst., vol. 334, no. 2, pp. 307-318, 1997, doi: https://doi.org/10.1016/S0016-0032(96)00063-4

[17] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” Ann. Math. Stat., vol. 22, no. 1, pp. 79-86, 1951, doi: https://doi.org/10.1214/aoms/1177729694

[18] P. A. Apellániz et al., “Synthetic Tabular Data Validation: A Divergence-Based Approach,” IEEE Access, vol. 12, pp. 103895-103907, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3434582

[19] X. Donath et al., “Next-generation sequencing identifies monogenic diabetes in 16% of patients with late adolescence/adult-onset diabetes selected on a clinical basis: a cross-sectional analysis,” BMC Med., vol. 17, no. 1, 2019, art. no. 132, doi: https://doi.org/10.1186/s12916-019-1363-0

[20] Y. Park, D. Heider, and A.-C. Hauschild, “Integrative Analysis of Next-Generation Sequencing for Next-Generation Cancer Research toward Artificial Intelligence,” Cancers, vol. 13, no. 13, 2021, art. no. 3148, doi: https://doi.org/10.3390/cancers13133148

[21] P. S. Paladugu et al., “Generative Adversarial Networks in Medicine: Important Considerations for this Emerging Innovation in Artificial Intelligence,” Ann. Biomed. Eng., vol. 51, pp. 2130-2142, 2023, doi: https://doi.org/10.1007/s10439-023-03304-z

[22] A. Aggarwal, M. Mittal, and G. Battineni, “Generative adversarial network: An overview of theory and applications,” Int. J. Inf. Manag. Data Insights, vol. 1, no. 1, 2021, art no. 100004, doi: https://doi.org/10.1016/j.jjimei.2020.100004

[23] C. O. S. Sorzano, J. Vargas, and A. Pascual, “A survey of dimensionality reduction techniques”, 2014, arXiv:1403.2877, doi: https://doi.org/10.48550/arXiv.1403.2877

[24] G. T. Reddy et al., “Analysis of Dimensionality Reduction Techniques on Big Data,” IEEE Access, vol. 8, pp. 54776-54788, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2980942

[25] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, “Missing Data in Clinical Research: A Tutorial on Multiple Imputation,” Can. J. Cardiol., vol. 37, no. 9, pp. 1322-1331, 2021, doi: https://doi.org/10.1016/j.cjca.2020.11.010

[26] T. Emmanuel et al., “A survey on missing data in machine learning,” J. Big Data, vol. 8, no. 1, 2021, art. no. 140, doi: https://doi.org/10.1186/s40537-021-00516-9

[27] R. S. Somasundaram and R. Nedunchezhian, “Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values,” Int. J. Comput. Appl., vol. 21, no. 10, pp. 14-19, 2011, doi: https://doi.org/10.5120/2619-3544

[28] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed., Hoboken, NJ, USA: Wiley, 2019, doi: https://doi.org/10.1002/9781119482260

[29] C. K. Enders, Applied Missing Data Analysis, 2nd ed., Guilford Publications, 2022.

[30] S. van Buuren, Flexible Imputation of Missing Data, 2nd ed., NY, USA: CRC Press, 2018, doi: https://doi.org/10.1201/9780429492259

[31] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed., O’Reilly Media, Inc., 2019.

[32] S. G. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage," Int. Adv. Res. J. Sci. Eng. Technol., pp. 20-22, 2015, doi: https://doi.org/10.17148/IARJSET.2015.2305

[33] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," Appl. Soft Comput., vol. 97, 2020, art. no. 105524, doi: https://doi.org/10.1016/j.asoc.2019.105524

[34] L. Rahmad Ramadhan and Y. Anne Mudya, "A Comparative Study of Z-Score and Min-Max Normalization for Rainfall Classification in Pekanbaru," J. Data Sci., vol. 2024, 2024, art. no. 4. [Online]. Available: https://iuojs.intimal.edu.my/index.php/jods/article/view/446

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, arXiv:7101.07875, doi: https://doi.org/10.48550/arXiv.1701.07875

[36] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems, 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[37] T. Salimans et al., "Improved Techniques for Training GANs," in Advances in Neural Information Processing Systems, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," 2016, arXiv:1511.06434, doi: https://doi.org/10.48550/arXiv.1511.06434

[39] S. Jenni and P. Favaro, "On Stabilizing Generative Adversarial Training With Noise," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 12137-12145, doi: https://doi.org/10.1109/CVPR.2019.01242

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 2010. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

[41] I. Gulrajani et al., "Improved Training of Wasserstein GANs," in NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295327

[42] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[43] J. Lin, "Divergence measures based on the Shannon entropy," IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 145-151, 1991, doi: https://doi.org/10.1109/18.61115

[44] M. Lucic et al., "Are GANs Created Equal? A Large-Scale Study," in NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 2018. [Online]. Available: https://dl.acm.org/doi/10.5555/3326943.3327008

[45] A. Borji, "Pros and cons of GAN evaluation measures," Comput. Vis. Image Underst., vol. 179, pp. 41-65, 2019, doi: https://doi.org/10.1016/j.cviu.2018.10.009

[46] S. Arora et al., "Generalization and Equilibrium in Generative Adversarial Nets (GANs)," in ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70, Sydney, Australia, 2017, pp. 224-232. [Online]. Available: https://dl.acm.org/doi/10.5555/3305381.3305405

[47] M. S. M. Sajjadi et al., "Assessing Generative Models via Precision and Recall," in NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 2018. [Online]. Available: https://dl.acm.org/doi/10.5555/3327345.3327429

[48] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization," in NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, España, 2016. [Online]. Available: https://dl.acm.org/doi/10.5555/3157096.3157127

[49] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001, doi: https://doi.org/10.1023/A:1010933404324

[50] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, pp. 18-22, 2002. [Online]. Available: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

[51] T. K. Ho, "Random decision forests," in Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 1995, pp. 278-282, doi: https://doi.org/10.1109/ICDAR.1995.598994

[52] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011. [Online]. Available: https://dl.acm.org/doi/10.5555/1953048.2078195

[53] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, vol. 6, no. 1, 2019, art. no. 60, doi: https://doi.org/10.1186/s40537-019-0197-0

[54] E. Choi et al., "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks," 2018, arXiv:1703.06490, doi: https://doi.org/10.48550/arXiv.1703.06490

[55] A. Goncalves et al., "Generation and evaluation of synthetic patient data," BMC Med. Res. Methodol., vol. 20, no. 1, 2020, art. no. 108, doi: https://doi.org/10.1186/s12874-020-00977-1

[56] L. Xu, et al., "Modeling Tabular data using Conditional GAN," 2019, arXiv:1907.00503, doi: https://doi.org/10.48550/arXiv.1907.00503